CHAPITRE 1

Statistique descriptive

La statistique descriptive a pour but de décrire, classer et simplifier des données qui peuvent être volumineuses, de les représenter de manière synthétique sous forme de tableaux ou de graphiques, et d'extraire quelques valeurs importantes qui décrivent les propriétés essentielles des données telles que la moyenne, la variance, la corrélation etc.

1.1 Vocabulaire

- **Population** L'ensemble sur lequel porte l'activité statistique s'appelle la population, généralement notée Ω. Lorsque la population est finie, le nombre d'éléments contenus dans Ω est noté N. Les éléments qui constituent la population sont appelés les individus ou encore les unités statistiques.
- **Échantillon** Un échantillon, noté généralement S (pour "*sample*") est une partie de la population prélevée soit de façon aléatoire soit de façon non aléatoire. Le nombre d'éléments de S est noté n.
- **Caractères** Les caractéristiques étudiées sur les individus d'une population sont appelées les caractères. Soit $\mathscr{C}$ l'ensemble des valeurs possibles du caractère, on définit alors un caractères comme une application $\chi : \Omega \to \mathscr{C}$ qui associe à chaque individu $\omega \in \Omega$ la valeur $\chi(\omega) \in \mathscr{C}$ que prend ce caractère sur l'individu ω.

  Il existe deux types de caractères :
  - caractères **quantitatifs** : c'est un caractère dont les issues produisent un nombre (caractères simples ou univariés, $\mathscr{C} \subset \mathbb{R}$) ou une suite de nombres (caractères multiples ou multivariés, $\mathscr{C} \subset \mathbb{R}^m$). Parmi les caractères quantitatifs il faut distinguer
    - les caractères quantitatifs **continus** qui peuvent prendre toutes les valeurs d'un intervalle,
    - les caractères quantitatifs **discrets** qui ne prennent que des valeurs isolées ;
  - caractères **qualitatifs** : c'est un caractère dont les issues ne sont pas quantifiables numériquement. On parle alors de modalités et non d'issues dans ce cas. Parmi les caractères qualitatifs il faut distinguer
    - les caractères qualitatifs **ordinaux** qui peuvent être ordonnées,
    - les caractères qualitatifs **nominaux**.

EXEMPLE
- La masse d'un individu est un caractère quantitatif univarié continu ($\mathscr{C} \subset \mathbb{R}^+$).
- Le relevé de températures d'une ville pendant le mois de juin est un caractère quantitatif multivarié continu ($\mathscr{C} \subset \mathbb{R}^{30}$).
- Le genre est un caractère qualitatif nominal ($\mathscr{C} = \{\text{homme}, \text{femme}\}$). On peut bien sûr coder la valeur "homme" par "0" et "femme" par "1" mais cela ne donne ni un sens à l'ordre ni le transforme en un caractère quantitatif.

1.1.1 Série statistique et distribution statistique non groupée

Considérons une série statistique associée à un caractère, c'est-à-dire un échantillon de n valeurs réelles $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$. Notons $\mathscr{C} = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ les valeurs atteintes par le caractère, *i.e.* $x_k \in \mathscr{C}$.
L'ordre dans lequel on a recueilli les données ne présentant pas d'intérêt particulier, on a intérêt à regrouper les données par paquets. On appelle alors

- **effectif** de la valeur α_i, et on le note n_i, le nombre de fois que la valeur $\alpha_i \in \mathscr{C}$ est prise dans $\mathbf{x}$;
- **effectif cumulé** en α_i la somme $\sum_{j=1}^{i} n_j$;
- **fréquence** de la valeur α_i le rapport $f_i = \frac{n_i}{n}$;
- **fréquence cumulée** en α_i la somme $\sum_{j=1}^{i} f_j$.

Si on écrit la **série statistique** $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme $(\alpha_i, n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_i)_{i\in[\![1;p]\!]}$ on parle de **distribution statistique**.

EXEMPLE
Soit la série statistique $\mathbf{x} = (1,1,2,1,1,0,3,1)$.

- Elle contient $n = 8$ valeurs $x_1 = 1$, $x_2 = 1$, $x_3 = 2$, $x_4 = 1$, $x_5 = 1$, $x_6 = 0$, $x_7 = 3$, $x_8 = 1$;
- les valeurs atteintes sont $\mathscr{C} = \{\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 2, \alpha_4 = 3\}$;
- $n_1 = 1$, $n_2 = 5$, $n_3 = 1$, $n_4 = 1$;
- $f_1 = 1/8$, $f_2 = 5/8$, $f_3 = 1/8$, $f_4 = 1/8$;
- les effectifs cumulés sont respectivement 1, 6, 7, 8 ;
- les fréquences cumulées sont respectivement $1/8$, $6/8$, $7/8$, $8/8 = 1$.

DATA 1.1 (ENFANTS)
Considérons le nombre d'enfants par famille collectés dans un immeuble de $n = 80$ familles :

$$\mathbf{x} = (0,3,0,0,0,0,3,3,3,5,3,2,0,0,0,1,2,1,1,1,1,1,2,0,4,2,2,0,4,1,0,5,2,3,2,3,0,3,4,5,0,1,3,0,0,3,$$
$$1,0,0,0,2,0,0,0,1,0,3,4,4,0,0,0,0,1,5,2,0,3,2,0,1,0,2,4,0,1,3,3,0,5).$$

On a $\mathscr{C} = \{0,1,2,3,4,5\}$.
L'effectif n_i de chaque valeur $\alpha_i \in \mathscr{C}$ est le nombre d'observations de cette valeurs (*i.e.* combien de fois α_i apparaît dans $\mathbf{x}$). La fréquence f_i de la valeur α_i est le rapport de l'effectif n_i sur le nombre totale d'observations n :

Valeur α_i (Nombre d'enfants)	Effectif n_i (Nombre de familles)	Fréquence f_i (Proportion de familles)	Effectif cumulé	Fréquence cumulée
0	31	31/80	31	31/80
1	13	13/80	44	44/80
2	11	11/80	55	55/80
3	14	14/80	69	69/80
4	6	6/80	75	75/80
5	5	5/80	80	80/80
	$\sum_{i=1}^{p=6} n_i = 80$	$\sum_{i=1}^{p=6} f_i = 1$		

1.1.2 Série statistique et distribution statistique groupée

Lorsqu'un caractère comprend un grand nombre de valeurs, il est préférable de les regrouper. L'ensemble $\mathscr{C}$ des valeurs du caractère est alors partagé en intervalles disjoints $]\alpha_i;\alpha_{i+1}]$, appelés **classes**, avec $\alpha_i < \alpha_{i+1}$.
On appelle alors

- **amplitude** de la classe $]\alpha_i;\alpha_{i+1}]$ la largeur de l'intervalle ;
- **effectif** de la classe $]\alpha_i;\alpha_{i+1}]$, et on le note n_i, le nombre de valeurs de $\mathbf{x}$ qui appartiennent à cet interval (le nombre d'observations qui tombent dans cette classe) ;
- **effectif cumulé** en α_i le nombre de valeurs de $\mathbf{x}$ qui appartiennent à $]-\infty;\alpha_i]$;
- **fréquence** de la classe $]\alpha_i;\alpha_{i+1}]$ le rapport $f_i = \frac{n_i}{n}$;
- **fréquence cumulée** en α_i la somme $\sum_{j=1}^{i} f_j$.

Si on écrit la série statistique $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme $(]\alpha_i;\alpha_{i+1}], n_i)_{i\in[\![1;p]\!]}$ ou $(]\alpha_i;\alpha_{i+1}], f_i)_{i\in[\![1;p]\!]}$ on parle de **distribution statistique groupée**.
Le nombre de classes ne doit pas être trop grand pour que le nouveau tableau soit suffisamment clair, mais pas trop petit pour qu'il n'y ait pas de perte d'information trop importante. Il faut enfin que toutes les observations soient recouvertes par ces classes.

DATA 1.2 (AMPOULES)
Supposons qu'on ait recueilli les durée de vie (en heures) d'un lot d'ampoules :

$$\mathbf{x} = (2560, 229323551738, 2272, 2259, 2549, 1688, 2306, 2494, 2131, 1864, 2107, 2056, 2557, 1311, 2305, 2433, 2408, 1523,$$
$$2155, 2531, 2327, 1396, 2414, 2411, 2329, 1424, 2456, 2149, 2039, 1447, 1884, 2289, 2340, 1428, 2134, 2333, 1989, 1554,$$
$$2558, 2031, 2111, 1415, 2335, 2546, 2343, 1493, 2435, 2131, 2026, 1631, 2513, 2233, 2416, 1441, 2475, 2304, 2177, 1432,$$
$$1918, 2092, 2139, 1657, 2628, 2334, 2091, 1428, 2504, 2519, 2125, 1458, 2085, 2234, 2339, 1484, 2052, 2168, 2280, 1547,$$
$$2393, 2048, 1517, 1579, 2373, 2207, 1452, 1859, 2177, 2112, 1573, 1473, 2474, 2513, 1488, 1391, 2109, 2296, 1410, 1607,$$

2286, 2303, 1432, 1577, 2389, 1945, 1589, 1438, 2408, 1925, 1431, 1652, 2215, 2420, 1546, 1597, 2429, 2381, 1672, 1636).

On peut regrouper ces données en 8 classes de même amplitudes :

Durée de vie	Effectif	Fréquence
]1200;1400]	3	3/120
]1400;1600]	27	27/120
]1600;1800]	8	8/120
]1800;2000]	7	7/120
]2000;2200]	23	23/120
]2200;2400]	28	28/120
]2400;2600]	23	23/120
]2600;2800]	1	1/120
	$\sum_{i=1}^{p=8} n_i = 120$	$\sum_{i=1}^{p=8} f_i = 1$

On peut aussi subdiviser les trois classes de 2000 à 2600 en six classes et obtenir ainsi des classes d'amplitudes différentes :

Durée de vie	Effectif	Fréquence
]1200;1400]	3	3/120
]1400;1600]	27	27/120
]1600;1800]	8	8/120
]1800;2000]	7	7/120
]2000;2100]	9	9/120
]2100;2200]	14	14/120
]2200;2300]	11	11/120
]2300;2400]	17	17/120
]2400;2500]	13	13/120
]2500;2600]	10	10/120
]2600;2800]	1	1/120
	$\sum_{i=1}^{p=11} n_i = 120$	$\sum_{i=1}^{p=11} f_i = 1$

Sur un caractère qualitatif, le seul calcul numérique qu'on puisse effectuer est le dénombrement des unités statistiques dans chaque catégorie de la variable qualitative.

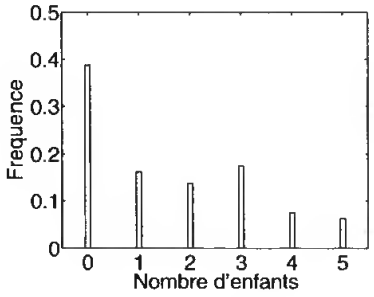
1.2 Données statistiques et leur représentation

1.2.1 Diagramme en bâtons

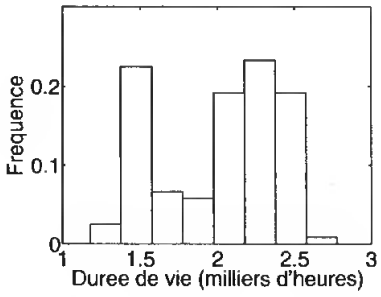
Dans le cas de données discrètes (ou d'une série statistique non groupée) on trace la plupart du temps un diagramme en bâtons des effectifs ou des fréquences. Des segments de droite verticaux sont dessinés. Chaque segment correspond à une classe (*i.e.* une modalité). La valeur de la classe est l'abscisse du segment, l'ordonnée de l'extrémité inférieure du segment est 0 et l'ordonnée de l'extrémité supérieure est l'effectif de la classe ou la fréquence. Les data 1.1 peut ainsi se représenter sous la forme du diagramme en bâtons donné sur la figure 1.1a.

1.2.2 Histogramme

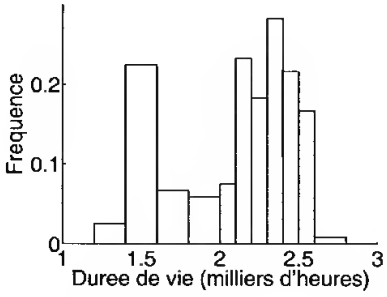
Dans le cas de données continues (ou d'une série statistique groupée), on regroupe d'abord les données par classes. On trace alors un histogramme constitué de rectangles verticaux. Les bases des rectangles sont déterminées par les classes. Les hauteurs de rectangles doivent être telles que les surfaces des rectangles sont proportionnelles aux effectifs des classes correspondantes.
Le travail est simple lorsque la largeur de chaque classe est la même : la hauteur d'un rectangle est alors prise égale à l'effectif (ou à la fréquence) de la classe correspondante. C'est le cas des data 1.2 avec la première subdivision en classes et on obtient alors la figure 1.1b.
Il arrive qu'on ait affaire à des classes non-régulières. Le tracé d'un histogramme doit alors prendre en compte la non-uniformité des largeurs des classes. Pour cela, on prend la plus petite des largeurs (ou amplitudes) comme largeur de référence, et multiplie la hauteur des rectangles par le rapport de leur largeur sur cette largeur minimale. C'est le cas de data 1.2 avec la deuxième subdivision en classes et on obtient alors la figure 1.1c.



(a) Distribution (diagramme en bâtons) du nombre d'enfants par famille.



(b) Histogramme des durées de vie des ampoules.



(c) Histogramme des durées de vie des ampoules.

FIGURE 1.1 – Exemples d'histogrammes.

1.3 Statistique descriptive univariée

Les tableaux et les diagrammes sont utiles, mais ils ne sont que des outils de visualisation. On cherche souvent, à partir de données quantitatives collectées, à extraire des caractéristiques chiffrées simples, des nombres qui révèlent les propriétés importantes de l'échantillon ou de la population. Nous nous intéressons à deux types de mesure : des mesures qui s'intéressent à la tendance centrale, *i.e.* à la plus représentative de toutes les données, et des mesures de la dispersion, *i.e.* combien les mesures de tendance centrale sont representatives de toutes les données.

1.3.1 Mesures de tendance centrale

Il s'agit de déterminer la valeur qui est la plus représentative de toutes les données.

Mode (ou classe modale)

Pour un caractère discret, le MODE est la valeur la plus fréquente que l'on trouve dans un échantillon.
Dans le cas de caractères continus ou plus généralement d'une série statistique groupée, on considère plutôt la CLASSE MODALE qui est le rapport fréquence/amplitude maximal. Le résultat dépend donc des classes choisies, ce qui rend cette notion peu pratique à utiliser.
Le mode n'est pas défini de manière unique. On peut trouver plusieurs classes avec le même effectif. On parle alors de distribution multi-modale.
Le mode est peu sensible aux valeurs extrêmes.

EXEMPLE

```
xx = [5 2 4 2 6]
mode(xx) % ans = 2

xx = [5 2 4 2 6 5]
mode(xx) % ans = 2
% If two, or more, values have the same frequency 'mode' returns the smallest
```

Médiane

La MÉDIANE est une valeur M telle qu'il y ait autant d'observations supérieures ou égales à M que d'observations inférieures ou égales à M. Pour calculer précisément la médiane, on commence par ordonner l'échantillon $\mathbf{x}$ par ordre croissant, et on note $\mathbf{y} = (y_k)_{k\in[\![1;n]\!]}$ l'échantillon ordonné tel que $y_1 \leq y_2 \leq \cdots \leq y_n$. Si l'échantillon comporte un nombre impair $2p+1$ d'observations, alors la médiane est

$$M(\mathbf{x}) = y_{p+1},$$

si l'échantillon est constitué d'un nombre pair $2p$ d'observations, alors la médiane est

$$M(\mathbf{x}) = \frac{y_p + y_{p+1}}{2}.$$

La médiane est peu sensible aux valeurs extrêmes et n'est pas forcément une modalité.

EXEMPLE
Si l'échantillon $\mathbf{x}$ est constitué de la suite d'entiers $(5,2,4,2,6)$, alors l'échantillon ordonné $\mathbf{y}$ est $(2,2,4,5,6)$. On a $n=5$ éléments, donc $p=\frac{n-1}{2}=2$ et $M(\mathbf{x})=y_{p+1}=y_3=4$.

```
xx = [5 2 4 2 6]
median(xx) % ans =4

xx = [5 2 4 2 6 30]
median(xx) % ans =4.5
```

Moyenne arithmétique

On peut définir la moyenne arithmétique d'une série statistique $\mathbf{x}=(x_k)_{k\in[\![1;n]\!]}$ comme étant le barycentre des données, affectées de coefficients égaux pour chaque individu :

$$\bar{\mathbf{x}}=\frac{1}{n}\sum_{k=1}^{n}x_k.$$

On appelle communément $\bar{\mathbf{x}}$ la moyenne de $\mathbf{x}$.
Si on écrit la série statistique $\mathbf{x}=(x_k)_{k\in[\![1;n]\!]}$ comme la distribution statistique $(\alpha_i,n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i,f_i)_{i\in[\![1;p]\!]}$, alors

$$\bar{\mathbf{x}}=\sum_{j=1}^{p}f_j\alpha_j=\frac{1}{n}\sum_{j=1}^{p}n_j\alpha_j.$$

Remarque (Sensibilité aux valeurs extrêmes)
La moyenne est très sensible aux valeurs extrêmes. Par exemple, si on cherche la fortune moyenne des Français à partir d'un échantillon de 1000 personnes, si l'une d'entre elles possède un milliard d'euros, alors la fortune moyenne est supérieure à un million d'euros quelles que soient les fortunes des 999 autres, puisqu'elle vérifie :

$$\bar{\mathbf{x}}=\frac{1}{10^3}\sum_{k=1}^{10^3}x_k\geq\frac{1}{10^3}\left(\sum_{k=1}^{999}0+10^9\right)=10^6.$$

EXEMPLE

```
xx = [5 2 4 2 6]
mean(xx) % ans = 3.8

xx = [5 2 4 2 6 50]
mean(xx) % ans = 11.5
```

Dans le cas d'une distribution statistique groupée $(]\alpha_i;\alpha_{i+1}],n_i)_{i\in[\![1;p]\!]}$ dont on n'a pas toutes les données $\mathbf{x}$, il n'est pas possible de calculer la moyenne exactement. Si on ne dispose que du tableau des fréquences, alors on estime la moyenne par la formule

$$\bar{\mathbf{x}}\approx\sum_{i=1}^{p}f_i\frac{\alpha_i+\alpha_{i+1}}{2},$$

où $\frac{\alpha_i+\alpha_{i+1}}{2}$ est le centre de la i-ème classe et f_i sa fréquence.

Propriété 1.1 (Fusion de données)
Considérons la situation où on dispose de deux échantillons $\mathbf{u}=(u_k)_{k\in[\![1;n_1]\!]}$ et $\mathbf{v}=(v_k)_{k\in[\![1;n_2]\!]}$ de tailles n_1 et n_2 et de moyennes respectives $\bar{\mathbf{u}}$ et $\bar{\mathbf{v}}$. L'échantillon globale $\mathbf{x}$ fusion des deux échantillons $\mathbf{u}$ et $\mathbf{v}$ est de taille $n=n_1+n_2$ et sa moyenne est

$$\bar{\mathbf{x}}=\frac{n_1\bar{\mathbf{u}}+n_2\bar{\mathbf{v}}}{n}.$$

Autrement dit, lorsqu'on fusionne les résultats issus d'échantillons différents, on peut obtenir la moyenne de l'échantillon global sans avoir à refaire tous les calculs.
La médiane et le mode ne vérifient pas cette propriété.

Propriété 1.2 (Erreur quadratique)
Considérons la fonction $\mathscr{E} : \mathbb{R} \to \mathbb{R}_+$ définie par

$$\mathscr{E}(\mu) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2.$$

Elle atteint son minimum en $\mu = \bar{\mathbf{x}}$.

La fonction $\mathscr{E}$, appelée ERREUR QUADRATIQUE, est la moyenne des carrés des distances entre les x_k et le nombre réel μ. La moyenne $\bar{\mathbf{x}}$ est la constante qui approche au mieux les données au sens des moindres carrés.

PREUVE
$\mathscr{E}(\mu) \geq 0$ pour tout μ et

$$\mathscr{E}'(\mu) = -\frac{2}{n}\sum_{k=1}^{n}(x_k - \mu) = -2\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \frac{1}{n}n\mu\right) = -2\left(\frac{1}{n}\sum_{k=1}^{n}x_k - \mu\right) = -2(\bar{\mathbf{x}} - \mu).$$

Ainsi $\mu = \bar{\mathbf{x}}$ est un extremum et comme $\mathscr{E}$ est quadratique, il s'agit d'un minimum.

1.3.2 Mesures de dispersion

On vient d'examiner différentes mesures de la tendance centrale d'un échantillon. On va maintenant chercher une mesure de la variabilité d'un échantillon, c'est-à-dire un nombre qui est d'autant plus grand que les données de l'échantillon sont dispersées.

Variance

La dispersion d'un échantillon peut se visualiser en considérant les écarts à la moyenne, c'est-à-dire l'échantillon $\mathbf{v} = (v_k)_{k\in[\![1;n]\!]}$ avec $v_k = x_k - \bar{\mathbf{x}}$. On cherche à obtenir une valeur unique représentative de ces écarts. On ne va pas prendre la moyenne $\bar{\mathbf{v}}$ car elle est nulle quel que soit l'échantillon (d'après la linéarité de la moyenne arithmétique). On va donc prendre les carrés des écarts et calculer leur moyenne arithmétique. On obtient ainsi la VARIANCE de l'échantillon :

$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2.$$

Autrement dit, la variance est la valeur de l'erreur quadratique en son minimum (la moyenne minimise la fonction "erreur quadratique", la variance est l'évaluation de cette fonction dans le minimum) :

$$V(\mathbf{x}) = \mathscr{E}(\bar{\mathbf{x}}).$$

La variance est une quantité positive qui augmente avec la dispersion des données. Elle est nulle si et seulement si toutes les données sont égales.
Si on écrit la série statistique $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme la distribution statistique $(\alpha_i, n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_i)_{i\in[\![1;p]\!]}$, alors

$$V(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{p} n_j(\alpha_j - \bar{\mathbf{x}})^2 = \sum_{j=1}^{p} f_j(\alpha_j - \bar{\mathbf{x}})^2.$$

Théorème 1.3 (de Koenig, formule de Huygens)

$$V(\mathbf{x}) = \frac{\sum_{k=1}^{n} x_k^2}{n} - \bar{\mathbf{x}}^2.$$

Cette expression est utile pour calculer pratiquement la variance d'un échantillon donné.

PREUVE

$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2. = \frac{1}{n}\sum_{k=1}^{n}(x_k^2 - 2x_k\bar{\mathbf{x}} + \bar{\mathbf{x}}^2). = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\frac{1}{n}\sum_{k=1}^{n}x_k\bar{\mathbf{x}} + \frac{1}{n}\sum_{k=1}^{n}\bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\bar{\mathbf{x}}^2 + \frac{1}{n}n\bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - \bar{\mathbf{x}}^2.$$

Si on écrit la série statistique $\mathbf{x} = (x_k)_{k\in[\![1;n]\!]}$ comme la distribution statistique $(\alpha_i, n_i)_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_i)_{i\in[\![1;p]\!]}$, alors la formule de Huygens devient

$$V(\mathbf{x}) = \frac{\sum_{j=1}^{p} n_j \alpha_j^2}{n} - \bar{\mathbf{x}}^2 = \sum_{j=1}^{p} f_j \alpha_j^2 - \bar{\mathbf{x}}^2.$$

Propriété 1.4 (Fusion de données)
Considérons la situation où on dispose de deux échantillons $\mathbf{u} = (u_k)_{k\in[\![1;n_1]\!]}$ et $\mathbf{v} = (v_k)_{k\in[\![1;n_2]\!]}$ de tailles n_1 et n_2 et de variances respectives $V(\mathbf{u})$ et $V(\mathbf{v})$. L'échantillon globale $\mathbf{x}$ fusion des deux échantillons $\mathbf{u}$ et $\mathbf{v}$ est de taille $n = n_1 + n_2$ et sa variance est

$$V(\mathbf{x}) = \frac{n_1}{n}\left(V(\mathbf{u}) + \bar{\mathbf{u}}^2\right) + \frac{n_2}{n}\left(V(\mathbf{v}) + \bar{\mathbf{v}}^2\right) - \left(\frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n}\right)^2.$$

PREUVE

$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{k=1}^{n_1} u_k^2 + \frac{1}{n}\sum_{k=1}^{n_2} v_k^2 - \left(\frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n_1 + n_2}\right)^2 = \frac{n_1}{n}\left(V(\mathbf{u}) + \bar{\mathbf{u}}^2\right) + \frac{n_2}{n}\left(V(\mathbf{v}) + \bar{\mathbf{v}}^2\right) - \left(\frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n}\right)^2.$$

Écart-type

La variance présente un inconvénient majeur : si les données s'expriment en unités physiques, la moyenne arithmétique s'exprime aussi dans cette unité, mais la variance s'exprime dans l'unité carrée. C'est pourquoi on a introduit la notion d'ÉCART-TYPE :

$$\sigma(\mathbf{x}) = \sqrt{V(\mathbf{x})}$$

Remarque (Diviser par n ou $n-1$?)
La variance est utilisée si la population est accessible dans sa totalité. Cependant d'ordinaire nous nous intéressons à une population dont on n'a pu mesurer qu'un échantillon. Dans ce cas, la meilleure estimation que l'on puisse faire de la variance de la population est

$$E(V) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2 = \frac{n}{n-1}V.$$

$E(V)$ est dit *variance corrigée* ou *estimateur non biaisé de la variance* de la population car si l'on multiplie le prélèvement d'échantillons de même effectif dans cette même population, la moyenne des $E(V)$ tend vers la variance réelle de la population, ce qui n'est pas le cas de V.
Même remarque pour l'*estimation de l'écart-type* de la population (ou *écart-type corrigé*)

$$s(\sigma) = \sqrt{E(V)} = \sqrt{\frac{n}{n-1}V} = \sqrt{\frac{n}{n-1}}\sigma$$

Faut-il calculer l'indice de dispersion de l'échantillon ou l'estimation de celui de la population, autrement dit, faut-il diviser par n ou $n-1$?

- Si l'on cherche à tirer des conclusions sur une population à partir d'un échantillon de celle-ci, $(n-1)$ sera utilisé dans les calculs,
- si l'on cherche à décrire un échantillon ou si la variable a été mesurée sur tous les individus de la population, c'est n qui sera utilisé.

EXEMPLE

```
xx = [5 2 4 2 6]
n=length(xx)

% Estimation de l'ecart-type
s=std(xx) % ans = 1.7889

% Estimation de la variance
V=var(xx) % ans = 3.2
% E=s*s

% Ecart-type
```

```
sigma=std(xx,1) % ans = 1.6
% sigma=sqrt((n-1)/n)*s

% Variance
var(xx,1) % ans = 2.56
% V=sigma*sigma
% V=(n-1)/n*E
```

EXEMPLE
Calculons le mode, la médiane, la moyenne arithmétique et la variance des data 1.1.

- **Mode** Le mode est 0.
- **Médiane** On a $n = 80$ éléments, donc $p = \frac{n}{2} = 40$ et $M(\mathbf{x}) = \frac{y_p + y_{p+1}}{2} = \frac{y_{40} + y_{41}}{2} = 1$ (car $y_i = 0$ pour $i = 1, \ldots, 31$, $y_i = 1$ pour $i = 32, \ldots, 44$ etc).
- **Moyenne** On a $\bar{\mathbf{x}} = \dfrac{0 \times 31 + 1 \times 13 + 2 \times 11 + 3 \times 14 + 4 \times 6 + 5 \times 5}{80} = 1.575$
- **Variance** On a $\dfrac{0^2 \times 31 + 1^2 \times 13 + 2^2 \times 11 + 3^2 \times 14 + 4^2 \times 6 + 5^2 \times 5}{80} = 5.05$ ainsi $V(\mathbf{x}) = 5.05 - 1.575 = 3.475$

```
clear all;

% Chargement des valeurs
load enfantsdata.dat ;
valeurs=sort(enfantsdata);

tt=0:5 ;
hist(valeurs,tt);
h = findobj(gca,'Type','patch');
set(h(1),'FaceColor','y','EdgeColor','k');

n=length(valeurs)
[effectif,c]=hist(valeurs,unique(valeurs))
freq=effectif/sum(effectif)

my_mode=valeurs(effectif==max(effectif))

my_moyenne=sum(c.*effectif)/sum(effectif)
my_moyenne=sum(c.*freq)
moyenne=mean(valeurs)

my_V=sum(valeurs.^2)/n-my_moyenne^2
V=var(valeurs,1)

my_sigma=sqrt(my_V)
sigma=std(valeurs,1)

mediane=median(valeurs)
figure()
pkg load statistics
boxplot(valeurs);
axis ([0,2,-1,6]);
```

Fractiles, quantiles

Les FRACTILES sont un autre moyen de quantifier la dispersion de données quantitatives. Le fractile à θ% d'un échantillon est la valeur qui sépare la fraction θ% des plus petites données de la fraction $(100 - \theta)$% des plus grandes données.
Le fractile à 50% n'est autre que la médiane.
Les fractiles à 25%, 50% et 75% sont les trois quartiles.
Une mesure de la dispersion d'un échantillon est l'ESPACE INTER-QUARTILE qui est la différence entre le troisième quartile et le premier quartile ; c'est donc la largeur de l'intervalle qui contient l'échantillon duquel on a retiré les 25% plus grandes valeurs et les 25% plis petites valeurs. Qualitativement, plus l'espace inter-quartile est grand, plus la dispersion des données est grande. L'espace inter-quartile est moins sensible aux valeurs extrêmes que l'écart-type.

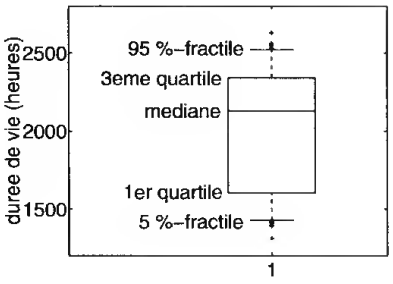


FIGURE 1.2 – Boîte à moustaches des durées de vie des ampoules (data 1.2).

1.3.3 Boîte à moustache

Une moyen très rapide de figurer le profil essentiel d'une série statistique quantitative est la boîte à moustaches (traduction française du terme "*Box and Whiskers Plot*" ou, en abrégé, "*Box Plot*"), aussi appelée boîte de distribution. Une telle boîte comprend

- une échelle de valeurs sur l'axe vertical ;
- le bord inférieur de la boîte correspond au premier quartile, noté Q_1 (*i.e.* le fractile à 25% ou quantile à 0.25) ;
- le bord supérieur de la boîte correspond au troisième quartile, noté Q_3 (*i.e.* le fractile à 75% ou quantile à 0.75) ;
- le trait horizontal au sein de la boîte correspond au deuxième quartile, noté Q_2 (*i.e.* la médiane) ;
- les moustaches inférieure et supérieure, représentées par des traits verticaux de chaque coté de la boîte et qui se terminent par des traits horizontaux (il existe plusieurs façon de construire les moustaches, parfois elles correspondent aux fractiles à 5% et 95%, parfois au premier et neuvième décile, mais d'autres conventions existent) ;
- les valeurs atypiques représentées par des cercles ou croix (on appelle ces données les outliers).

Une boîte avec des moustaches courtes indique que l'échantillon est assez dispersé.
Les boîtes à moustaches sont des résumés graphiques efficaces des données et sont donc très utiles pour comparer des distributions d'un groupe à l'autre. Contrairement à un histogramme, elle ne nécessitent pas de regrouper les observations en classes, ce qui est un avantage car le choix des classes est une opération subjective et qui influence fortement l'allure de l'histogramme construit à partir de celles-ci.
Sur la figure 1.2 on dessine la boîte à moustaches correspondant aux data 1.2. On constate que l'échantillon n'est pas équilibré, la médiane n'est pas vraiment au milieu des premier et troisième quartiles.

EXEMPLE
Utilisons les data 1.2.

```
clear all

% Chargement des valeurs
load mesures.dat
valeurs=sort(mesures);

% Nombre d'elements
n=length(valeurs)

%Moyenne
moyenne=mean(valeurs)

%Mediane
mediane=median(valeurs)

% Estimation de l'ecart-type sans biais
etsb=std(valeurs)

% Incertitude de type A
ua=etsb/sqrt(n)

% Definitions des bornes des intervalles
tt=1200:200:2800 ;
hist(valeurs,tt-100); # on passe le centre des intervalles
h = findobj(gca,'Type','patch');
display(h)
set(h(1),'FaceColor','r','EdgeColor','k');
```

```
% Decoupage en classes
ncl=length(tt)-1

%calcul des frequence
for i=1:ncl
  eff(i)=length(find(valeurs>tt(i) & valeurs<=tt(i+1))) ;
  fm(i)=eff(i)/n ;
end

% Borne gauche, borne droite, effectif, frequence
A=[tt(1:end-1)',tt(2:end)',eff',fm']

% Boite a moustache
figure()
pkg load statistics
boxplot(valeurs)
axis ([0,2]);

% The returned matrix has one column for each data set as follows:
% 1 Minimum = 1311.0
% 2 1st quartile = 1604.5
% 3 2nd quartile (median) = 2131.0
% 4 3rd quartile = 2340.8
% 5 Maximum = 2628.0
% 6 Lower confidence limit for median = 2025.5
% 7 Upper confidence limit for median = 2236.5
```

1.4 Statistique descriptive à deux caractères

Lorsque les observations portent simultanément sur deux caractères, on les présente sous la forme d'un tableau à double entrée. On définit alors la distribution conjointe, les distributions marginales et les distributions conditionnelles. L'étude de la distribution de deux variables se poursuit par celle de leur liaison. L'étude de la liaison entre les variables observées, appelée communément l'étude des corrélations, dépend de leur nature. Ici on n'envisagera que le cas de deux variables quantitatives non groupée en classes.

1.4.1 Distribution conjointe

Considérons donc une série statistique dont les observations portent sur deux caractères. On veut ici extraire des informations sur la distribution jointe des deux caractères et étudier leur dépendance. Désignons par

- $(\mathbf{x},\mathbf{y}) = (x_k, y_k)_{k\in[\![1;n]\!]}$ les n données brutes, généralement présentées sous la forme d'un tableau à deux colonnes ;
- $\mathscr{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_p\}$ les p modalités de $\mathbf{x}$, *i.e.* les p valeurs distinctes observées pour $\mathbf{x}$ (autrement dit $x_k \in \mathscr{A}$) ;
- $\mathscr{B} = \{\beta_1, \beta_2, \ldots, \beta_q\}$ les q modalités de $\mathbf{y}$, *i.e.* les q valeurs distinctes observées pour $\mathbf{y}$ (autrement dit $y_k \in \mathscr{B}$).

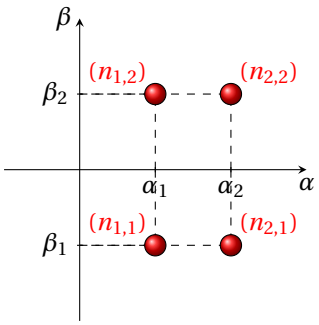
La répartition des n observations, ou **distribution conjointe**, suivant les modalités de $\mathbf{x}$ et $\mathbf{y}$ se présente sous forme d'un tableau à double entrée, appelée **tableau de contingence** :

Modalités de $\mathbf{y}$ / Modalités de $\mathbf{x}$	β_1	…	β_j	…	β_q	Effectif marginal de α_i
α_1	$n_{1,1}$	…	$n_{1,j}$	…	$n_{1,q}$	$n_{1,\cdot} = \sum_{j=1}^{q} n_{1,j}$
⋮	⋮		⋮		⋮	⋮
α_i	$n_{i,1}$	…	$n_{i,j}$	…	$n_{i,q}$	$n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j}$
⋮	⋮		⋮		⋮	⋮
α_p	$n_{p,1}$	…	$n_{p,j}$	…	$n_{p,q}$	$n_{p,\cdot} = \sum_{j=1}^{q} n_{p,j}$
Effectif marginal de β_j	$n_{\cdot,1} = \sum_{i=1}^{p} n_{i,1}$	…	$n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j}$	…	$n_{\cdot,q} = \sum_{i=1}^{p} n_{i,q}$	$n = \sum_{j=1}^{q} n_{\cdot,j} = \sum_{i=1}^{p} n_{i,\cdot}$

On appelle

- **effectif du couple** (α_i, β_j), et on le note $n_{i,j}$, le nombre de fois où le couple (α_i, β_j) est pris (*i.e.* le nombre de fois où la modalité α_i et la modalité β_j ont été observées simultanément) ;
- **fréquence du couple** (α_i, β_j) le rapport $f_{i,j} = \frac{n_{i,j}}{n}$.

Si on écrit la série statistique $(x_k, y_k)_{k\in[\![1;n]\!]}$ comme $((\alpha_i, \beta_j), n_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$ ou $((\alpha_i, \beta_j), f_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$ on parle de distribution conjointe.
On peut bien sûr représenter la série statistique ou la distribution conjointe sur un plan comme un nuage de points : chaque point correspond à un couple (α_i, β_j) affecté de son poids $n_{i,j}$, autrement dit chaque point correspond à une observation (x_k, y_k) et à coté on indique combien de fois cette observation apparaît. Il y aura donc $p \times q$ points (autant que de cases que dans le tableau de contingence), chaque point se trouvant sur un coin de la grille de coordonnées (α_i, β_j). Si pour un couple on a $n_{i,j} = 0$, on n'affichera pas de point. Si $n_{i,j} = 1$ pour tout $i = 1, \dots p$ et $j = 1, \dots q$, on a le nuage de points classique vu au chapitre précédent.



1.4.2 Distributions marginales

On peut bien sûr mener une étude statistique de chacun des caractères séparément, *i.e.* calculer la moyenne et la variance de chacune des séries simples ($\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, $V(\mathbf{x})$, $V(\mathbf{y})$). On appelle

- **effectif marginal** de α_i, et on le note $n_{i,\cdot}$, le nombre total d'observations de la modalité α_i de $\mathbf{x}$ quelle que soit la modalité de $\mathbf{y}$:

$$n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j};$$

- **effectif marginal** de β_j, et on le note $n_{\cdot,j}$, total d'observations de la modalité β_j de $\mathbf{y}$ quelle que soit la modalité de $\mathbf{x}$:

$$n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j};$$

- **fréquence marginale** de α_i le rapport $f_{i,\cdot} = \frac{n_{i,\cdot}}{n} = \sum_{j=1}^{q} f_{i,j}$;
- **fréquence marginale** de β_j le rapport $f_{\cdot,j} = \frac{n_{\cdot,j}}{n} = \sum_{i=1}^{p} f_{i,j}$.

On a bien évidemment

$$\sum_{i=1}^{p} n_{i,\cdot} = \sum_{j=1}^{q} n_{\cdot,j} = n \qquad \sum_{i=1}^{p} f_{i,\cdot} = \sum_{j=1}^{q} f_{\cdot,j} = 1.$$

Si on écrit la série statistique $\mathbf{x}$ comme $(\alpha_i, n_{i,\cdot})_{i\in[\![1;p]\!]}$ ou $(\alpha_i, f_{i,\cdot})_{i\in[\![1;p]\!]}$ on parle de distribution marginale de $\mathbf{x}$; de la même manière si on écrit la série statistique $\mathbf{y}$ comme $(\beta_j, n_{\cdot,j})_{j\in[\![1;q]\!]}$ ou $(\beta_j, f_{\cdot,j})_{j\in[\![1;q]\!]}$ on parle de distribution marginale de $\mathbf{y}$.
Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable en représentant graphiquement sa distribution et s'il s'agit d'une variable quantitative, en calculant ses caractéristiques de tendance centrale, de dispersion, de forme…
On appelle

- **moyenne marginale** de $\mathbf{x}$ la quantité

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}\alpha_i = \sum_{i=1}^{p} f_{i,\cdot}\alpha_i$$

- **moyenne marginale** de $\mathbf{y}$ la quantité

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}\beta_j = \sum_{j=1}^{q} f_{\cdot,j}\beta_j$$

• **variance marginale** de **x** la quantité

$$V(\mathbf{x}) = \frac{\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})^2}{n} = \frac{\sum_{k=1}^{n} x_k^2}{n} - \bar{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}(\alpha_i)^2 - \bar{\mathbf{x}}^2 = \sum_{i=1}^{p} f_{i,\cdot}\alpha_i^2 - \bar{\mathbf{x}}^2$$

• **variance marginale** de **y** la quantité

$$V(\mathbf{y}) = \frac{\sum_{k=1}^{n}(y_k - \bar{\mathbf{y}})^2}{n} = \frac{\sum_{k=1}^{n} y_k^2}{n} - \bar{\mathbf{y}}^2 = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}(\beta_j)^2 - \bar{\mathbf{y}}^2 = \sum_{j=1}^{q} f_{\cdot,j}\beta_j^2 - \bar{\mathbf{y}}^2.$$

1.4.3 Distributions conditionnelles

Une distribution à deux caractères présente deux types de distributions conditionnelles : les distributions conditionnelles de **x** selon **y** et les distributions conditionnelles de **y** selon **x**.

• **Distributions conditionnelles de y selon x**

Considérons la sous-population correspondante aux individus tels que $x = \alpha_i$.

La distribution de la variable **y** sachant $x = \alpha_i$ est appelée distribution conditionnelle de **y** pour $x = \alpha_i$. Il existe p distributions conditionnelle de **y** sachant $x = \alpha_i$, car $i = 1, \ldots, p$.

Modalités de **y** sachant α_i	β_1	…	β_j	…	β_q	Effectif marginal de α_i
α_i	$n_{i,1}$	…	$n_{i,j}$	…	$n_{i,q}$	$n_{i,\cdot} = \sum_{j=1}^{q} n_{i,j}$

Chaque distribution contient $n_{i,\cdot}$ observations et on peut calculer les quantités conditionnelles suivantes :

• **fréquence conditionnelle** de β_j sachant α_i comme la quantité

$$f_{j|i} = \frac{n_{i,j}}{n_{i,\cdot}} = \frac{f_{i,j}}{f_{i,\cdot}} \text{ avec } \sum_{j=1}^{q} f_{j|i} = 1;$$

• **distribution conditionnelle des fréquences** de **y** sachant α_i la distribution $(\beta_j, f_{j|i})_{j\in[\![1;q]\!]}$;

• **moyenne conditionnelle** de **y** sachant α_i la quantité

$$\bar{\mathbf{y}}\big|_i = \frac{1}{n_{i,\cdot}}\sum_{j=1}^{q} n_{i,j}\beta_j = \sum_{j=1}^{q} f_{j|i}\beta_j;$$

• **variance conditionnelle** de **y** sachant α_i la quantité

$$V_i(\mathbf{y}) = \frac{1}{n_{i,\cdot}}\sum_{j=1}^{q} n_{i,j}\beta_j^2 - \bar{\mathbf{y}}\big|_i^2 = \sum_{j=1}^{q} f_{j|i}\beta_j^2 - \bar{\mathbf{y}}\big|_i^2.$$

Les p modalités de **x** induisant une partition des n observations en p sous-groupes, la moyenne $\bar{\mathbf{y}}$ peut s'exprimer comme somme pondérées des p moyennes $\bar{\mathbf{y}}\big|_i$:

$$\bar{\mathbf{y}} = \sum_{i=1}^{p} \bar{\mathbf{y}}\big|_i f_{i,\cdot}$$

Il est fréquent de présenter les fréquences conditionnelles $f_{j|i}$ de **y** dans un tableau dont toutes les sommes en ligne

sont égales à 1 ; ce tableau est appelé tableau des profils en ligne :

Modalités de **y** / Modalités de **x**	β_1	…	β_j	…	β_q	
α_1	$f_{1\|1}$	…	$f_{j\|1}$	…	$f_{q\|1}$	1
⋮	⋮		⋮		⋮	⋮
α_i	$f_{1\|i}$	…	$f_{j\|i}$	…	$f_{q\|i}$	1
⋮	⋮		⋮		⋮	⋮
α_p	$f_{1\|p}$	…	$f_{j\|p}$	…	$f_{q\|p}$	1
Fréquence marginale de β_j	$f_{\cdot,1}$	…	$f_{\cdot,j}$	…	$f_{\cdot,q}$	1

- **Distributions conditionnelles de x selon y**

  De manière analogue, considérons maintenant la sous-population correspondante aux individus tels que $y = \beta_j$. La distribution de la variable **x** sachant $y = \beta_j$ est appelée distribution conditionnelle de **x** pour $y = \beta_j$. Il existe q distributions conditionnelle de **x** sachant $y = \beta_j$ car $j = 1, \ldots, q$.

Modalités de **x** sachant β_j	β_j
α_1	$n_{1,j}$
⋮	⋮
α_i	$n_{i,j}$
⋮	⋮
α_p	$n_{p,j}$
Effectif marginal de β_j	$n_{\cdot,j} = \sum_{i=1}^{p} n_{i,j}$

  Chaque distribution contient $n_{\cdot,j}$ observations et on peut définir les quantités conditionnelles suivantes :

  - **fréquence conditionnelle** de α_i sachant β_j comme la quantité

$$f_{i|j} = \frac{n_{i,j}}{n_{\cdot,j}} = \frac{f_{i,j}}{f_{\cdot,j}}$$

  - **distribution conditionnelle des fréquences** de **x** sachant β_j la distribution $(\alpha_i, f_{i|j})_{i \in [\![1;p]\!]}$;
  - **moyenne conditionnelle** de **x** sachant β_j la quantité

$$\bar{\mathbf{x}}|_j = \frac{1}{n_{\cdot,j}} \sum_{i=1}^{p} n_{i,j} \alpha_i = \sum_{i=1}^{p} f_{j|i} \alpha_i$$

  - **variance conditionnelle** de **x** sachant β_j la quantité

$$V_j(\mathbf{x}) = \frac{1}{n_{\cdot,j}} \sum_{i=1}^{p} n_{i,j} \alpha_i^2 - \bar{\mathbf{x}}|_j^2 = \sum_{i=1}^{p} f_{i|j} \alpha_i^2 - \bar{\mathbf{x}}|_j^2 .$$

  Les q modalités de **y** induisant une partition des n observations en q sous-groupes, la moyenne $\bar{\mathbf{x}}$ peut s'exprimer comme somme pondérées des q moyennes $\bar{\mathbf{x}}|_j$:

$$\bar{\mathbf{x}} = \sum_{j=1}^{q} \bar{\mathbf{x}}|_j \, f_{\cdot,j}$$

  De même on présente les fréquences conditionnelles $f_{i|j}$ de **x** dans un tableau dont toutes les sommes en colonne

sont égales à 1 ; ce tableau est appelé tableau des profils en colonne :

Modalités de **y** / Modalités de **x**	β_1	…	β_j	…	β_q	Fréquence marginale de α_i
α_1	$f_{1\|1}$	…	$f_{1\|j}$	…	$f_{1\|q}$	$f_{1,\cdot}$
⋮	⋮		⋮		⋮	⋮
α_i	$f_{i\|1}$	…	$f_{i\|j}$	…	$f_{i\|q}$	$f_{i,\cdot}$
⋮	⋮		⋮		⋮	⋮
α_p	$f_{p\|1}$	…	$f_{p\|j}$	…	$f_{p\|q}$	$f_{p,\cdot}$
	1	…	1	…	1	1

EXEMPLE

Soient les données brutes $\big((1,0),(1,2),(2,0),(2,2),(2,2),(1,1)\big)$, alors $n=6$.

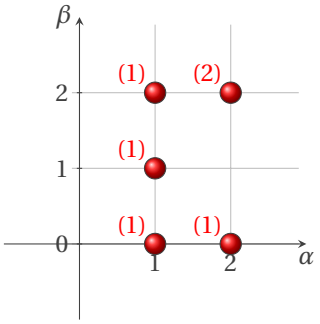
On a $x_k \in \mathscr{A} = \{1,2\}$ et $y_k \in \mathscr{B} = \{0,1,2\}$ pour tout $k = 1,2,\ldots,n$, ainsi $p=2$ et $q=3$. Écrivons les observations dans un tableau à deux colonnes :

x	**y**
1	0
1	2
2	0
2	2
2	2
1	1

- **Distribution conjointe et distributions marginales**

Le tableau des contingences avec les effectifs de chaque couple et les effectifs marginaux est

$\mathscr{A}$ / $\mathscr{B}$	$\beta_1=0$	$\beta_2=1$	$\beta_3=2$	Effectif marginal de α_i
$\alpha_1=1$	$n_{1,1}=1$	$n_{1,2}=1$	$n_{1,3}=1$	$n_{1,\cdot}=3$
$\alpha_2=2$	$n_{2,1}=1$	$n_{2,2}=0$	$n_{2,3}=2$	$n_{2,\cdot}=3$
Effectif marginal de β_j	$n_{\cdot,1}=2$	$n_{\cdot,2}=1$	$n_{\cdot,3}=3$	$n=6$



Le tableau des contingences avec les fréquences de chaque couple et les fréquences marginales est

$\mathscr{A}$ / $\mathscr{B}$	$\beta_1=0$	$\beta_2=1$	$\beta_3=2$	Fréquence marginale de α_i
$\alpha_1=1$	$f_{1,1}=1/6$	$f_{1,2}=1/6$	$f_{1,3}=1/6$	$f_{1,\cdot}=3/6$
$\alpha_2=2$	$f_{2,1}=1/6$	$f_{2,2}=0/6$	$f_{2,3}=2/6$	$f_{2,\cdot}=3/6$
Fréquence marginale de β_j	$f_{\cdot,1}=2/6$	$f_{\cdot,2}=1/6$	$f_{\cdot,3}=3/6$	1

Les moyennes marginales de **x** et **y** sont

$$\bar{\mathbf{x}} = \frac{1}{n}\left(n_{1,\cdot}\alpha_1 + n_{2,\cdot}\alpha_2\right) = \frac{1}{6}\left(3\alpha_1 + 3\alpha_2\right) = \frac{3}{2},$$

$$\bar{\mathbf{y}} = \frac{1}{n}\left(n_{\cdot,1}\beta_1 + n_{\cdot,2}\beta_2 + n_{\cdot,3}\beta_3\right) = \frac{1}{6}\left(2\beta_1 + 1\beta_2 + 3\beta_3\right) = \frac{7}{6}.$$

- **Distributions conditionnelles y sachant x**

  - **y sachant α_1** On ne considère que la ligne de la modalité α_1 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	Fréquence marginale de α_1
$\alpha_1 = 1$	1/6	1/6	1/6	$f_{1,\cdot} = 3/6$

$$f_{j=1|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_1$$
$$f_{j=2|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_1$$
$$f_{j=3|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_3 \text{ sachant } \alpha_1$$

De plus,

$$\sum_{j=1}^{q=3} f_{j|i=1} = 1$$
$$\bar{\mathbf{y}}\big|_{i=1} = \sum_{j=1}^{q=3} f_{j|i=1}\beta_j = \frac{1}{3}\beta_1 + \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 = 1 \qquad \text{moyenne conditionnelle de } \mathbf{y} \text{ sachant } \alpha_1$$

  - **y sachant α_2** On ne considère que la ligne de la modalité α_2 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	Fréquence marginale de α_2
$\alpha_2 = 2$	1/6	0/6	2/6	$f_{2,\cdot} = 3/6$

$$f_{j=1|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_2$$
$$f_{j=2|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{0/6}{3/6} = 0 \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_2$$
$$f_{j=3|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{2/6}{3/6} = \frac{2}{3} \qquad \text{fréquence conditionnelle de } \beta_3 \text{ sachant } \alpha_2$$

De plus,

$$\sum_{j=1}^{q=3} f_{j|i=2} = 1$$
$$\bar{\mathbf{y}}\big|_{i=2} = \sum_{j=1}^{q=3} f_{j|i=2}\beta_j = \frac{1}{3}\beta_1 + 0\beta_2 + \frac{2}{3}\beta_3 = \frac{4}{3} \qquad \text{moyenne conditionnelle de } \mathbf{y} \text{ sachant } \alpha_2$$

  - **tableau des profils en ligne** $f_{j|i}$

Modalités de **y** / Modalités de **x**	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	
$\alpha_1 = 1$	$f_{1\|1} = 1/3$	$f_{2\|1} = 1/3$	$f_{3\|1} = 1/3$	1
$\alpha_2 = 2$	$f_{1\|2} = 1/3$	$f_{2\|2} = 0$	$f_{3\|2} = 2/3$	1
Fréquence marginale de β_j	$f_{\cdot,1} = 2/6$	$f_{\cdot,2} = 1/6$	$f_{\cdot,3} = 3/6$	1

On a bien

$$\sum_{i=1}^{p=2} \bar{\mathbf{y}}|_i \, f_{i,\cdot} = 1\frac{3}{6} + \frac{4}{3}\frac{3}{6} = \frac{7}{6} = \bar{\mathbf{y}},$$

• **Distributions conditionnelles x sachant y**

• **x sachant** β_1 On ne considère que la colonne de la modalité β_1 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$
$\alpha_1 = 1$	1/6
$\alpha_2 = 2$	1/6
Fréquence marginale de β_1	$f_{\cdot,1} = 2/6$

$$f_{i=1|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{2/6} = \frac{1}{2} \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_1$$

$$f_{i=2|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{2/6} = \frac{1}{2} \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_1$$

De plus,

$$\sum_{i=1}^{p=2} f_{i|j=1} = 1$$

$$\bar{\mathbf{x}}|_{j=1} = \sum_{i=1}^{p=2} f_{i|j=1}\alpha_i = \frac{1}{2}\alpha_1 + \frac{1}{2}\alpha_2 = \frac{3}{2} \qquad \text{moyenne conditionnelle de } \mathbf{x} \text{ sachant } \beta_1$$

• **x sachant** β_2 On ne considère que la colonne de la modalité β_2 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_2 = 1$
$\alpha_1 = 1$	1/6
$\alpha_2 = 2$	0/6
Fréquence marginale de β_2	$f_{\cdot,2} = 1/6$

$$f_{i=1|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{1/6} = 1 \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_2$$

$$f_{i=2|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{0/6}{1/6} = 0 \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_2$$

De plus,

$$\sum_{i=1}^{p=2} f_{i|j=2} = 1$$

$$\bar{\mathbf{x}}|_{j=2} = \sum_{i=1}^{p=2} f_{i|j=2}\alpha_i = 1\alpha_1 + 0\alpha_2 = 1 \qquad \text{moyenne conditionnelle de } \mathbf{x} \text{ sachant } \beta_2$$

• **x sachant** β_3 On ne considère que la colonne de la modalité β_3 :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_3 = 2$
$\alpha_1 = 1$	1/6
$\alpha_2 = 2$	2/6
Fréquence marginale de β_3	$f_{\cdot,3} = 3/6$

$$f_{i=1|j=3} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{1/6}{3/6} = \frac{1}{3} \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_3$$

$$f_{i=2|j=3} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{2/6}{3/6} = \frac{2}{3} \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_3$$

De plus,

$$\sum_{i=1}^{p=3} f_{i|j=3} = 1$$

$$\bar{\mathbf{x}}|_{j=3} = \sum_{i=1}^{p=2} f_{i|j=3}\alpha_i = \frac{1}{3}\alpha_1 + \frac{2}{3}\alpha_2 = \frac{5}{3} \qquad \text{moyenne conditionnelle de } \mathbf{x} \text{ sachant } \beta_3.$$

- **tableau des profils en colonne** $f_{i|j}$

Modalités de **y** / Modalités de **x**	$\beta_1 = 0$	$\beta_2 = 1$	$\beta_3 = 2$	Fréquence marginale de α_i
$\alpha_1 = 1$	$f_{1\|1} = 1/2$	$f_{1\|2} = 1$	$f_{1\|3} = 1/3$	$f_{1,\cdot}$
$\alpha_2 = 2$	$f_{2\|1} = 1/2$	$f_{2\|2} = 0$	$f_{2\|3} = 2/3$	$f_{p,\cdot}$
	1	1	1	1

On a bien

$$\sum_{j=1}^{q=3} \bar{\mathbf{x}}|_j \, f_{\cdot,j} = \frac{3}{2}\frac{2}{6} + 1\frac{1}{6} + \frac{5}{3}\frac{3}{6} = \frac{3}{2} = \bar{\mathbf{x}}.$$

1.4.4 Indépendance statistique

Si tous les profils en colonne du tableau en colonne sont identiques, cela signifie que la distribution de la variable **x** ne dépend pas de la variable **y**, on dit alors que les variables **x** et **y** sont statistiquement indépendantes dans l'ensemble des n individus considérés, et dans ce cas toutes les distributions conditionnelles de **x** sont identiques à la distribution marginale de **x**. Par raison de symétrie, l'indépendance statistique entre **x** et **y** implique aussi des profils en ligne identiques à la distribution marginale de **y**.
Les deux séries **x** et **y** sont indépendantes si et seulement si

$$\begin{cases} f_{i|j} = f_{i,\cdot}, & \textit{i.e.} \text{ la distribution conditionnelle des fréquences de } \alpha_i \text{ sachant } \beta_j \text{ ne dépend pas de } j \\ f_{j|i} = f_{\cdot,j}, & \textit{i.e.} \text{ la distribution conditionnelle des fréquences de } \beta_j \text{ sachant } \alpha_i \text{ ne dépend pas de } i. \end{cases}$$

De plus, si les deux séries sont indépendantes, alors pour tout $i = 1,\dots,p$ et $j = 1,\dots,q$

$$f_{i,j} = f_{i,\cdot} f_{\cdot,j}$$

Lorsque deux variables dépendent statistiquement l'une de l'autre, on cherche à évaluer l'intensité de leur liaison et dans le cas de deux variables quantitatives, on examine si on peut les considérer liées par une relation linéaire.

1.4.5 Liaison entre deux variables quantitatives : covariance et corrélation

La dispersion d'une série bivarié $(x_k, y_k)_{k\in[\![1;n]\!]}$ peut se visualiser en considérant les écarts aux deux moyennes. On cherche à obtenir une valeur unique représentative de ces écarts. On obtient ainsi la **covariance** de la série $(x_k, y_k)_{k\in[\![1;n]\!]}$:

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})(y_k - \bar{\mathbf{y}}).$$

Si on écrit la série comme la distribution $((\alpha_i,\beta_j), n_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$ ou $((\alpha_i,\beta_j), f_{i,j})_{\substack{i\in[\![1;p]\!]\\ j\in[\![1;q]\!]}}$, on a

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j}(\alpha_i - \bar{\mathbf{x}})(\beta_j - \bar{\mathbf{y}}) = \sum_{j=1}^{q} f_{i,j}(\alpha_i - \bar{\mathbf{x}})(\beta_j - \bar{\mathbf{y}}).$$

Propriété 1.5

1. $C(\mathbf{x},\mathbf{y}) = C(\mathbf{y},\mathbf{x})$
2. $C(\mathbf{x},\mathbf{x}) = V(\mathbf{x})$ et $C(\mathbf{y},\mathbf{y}) = V(\mathbf{y})$
3. $V(\mathbf{x}+\mathbf{y}) = V(\mathbf{x}) + 2C(\mathbf{x},\mathbf{y}) + V(\mathbf{y})$
4. $C(a\mathbf{x}+b, c\mathbf{y}+d) = acC(\mathbf{x},\mathbf{y})$ pour tout $a,b,c,\in\mathbb{R}$
5. $C(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}\sum_{j=1}^{q}\alpha_i\beta_j f_{ij} - \bar{\mathbf{x}}\bar{\mathbf{y}}$
6. $|C(\mathbf{x},\mathbf{y})| = \sqrt{V(\mathbf{x})V(\mathbf{y})}$

Si **x** et **y** sont indépendantes alors la covariance est nulle. La réciproque est fausse : en effet la covariance mesure uniquement la dépendance linéaire.

Remarque (Diviser par n ou $n-1$?)
Dans la définition ci-dessus, le dénominateur est n. Si l'on tente d'estimer la covariance de la population à partir d'un échantillon il faudra diviser par $(n-1)$. Les notations de la covariance de l'échantillon et de l'estimation de celle de la population ne sont pas en générale distinguables. Ainsi, lorsqu'on utilise un logiciel, toujours faire un calcul d'essai pour connaître la formule utilisée. Dans Octave ou Matlab c'est $(n-1)$ qui est utilisé par défaut, mais on peut forcer l'utilisation de n, comme on voit dans l'exemple ci-dessous.
Dans ce chapitre on utilisera la notation

$$E(C(\mathbf{x},\mathbf{y})) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})(y_k-\bar{\mathbf{y}}) = \frac{n-1}{n}C(\mathbf{x},\mathbf{y}).$$

Comme pour la variance, on dispose d'une formule alternative pour la covariance qu'on utilise en pratique pour calculer une covariance :

Propriété 1.6

$$C(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{n}x_k y_k}{n} - \bar{\mathbf{x}}\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}\alpha_i\beta_j - \bar{\mathbf{x}}\bar{\mathbf{y}}.$$

EXEMPLE
Considérons l'échantillon bivarié $((1,1),(2,3),(3,5))$. On a

$$\mathbf{x} = (1,2,3) \qquad \mathbf{y} = (1,3,5) \qquad \bar{\mathbf{x}} = 2 \qquad \bar{\mathbf{y}} = 3$$

ainsi

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})(y_k-\bar{\mathbf{y}}) = \frac{(1-2)(1-3)+(2-2)(3-3)+(3-2)(5-3)}{3} = \frac{4}{3}$$

tandis que

$$E(C(\mathbf{x},\mathbf{y})) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})(y_k-\bar{\mathbf{y}}) = \frac{(1-2)(1-3)+(2-2)(3-3)+(3-2)(5-3)}{2} = \frac{4}{2} = 2.$$

'Octave'

```
x = [1 2 3];
y = [1 3 5];
E_cov = cov(x,y) % ans = 2
Cov = cov(x,y,1) % ans = 1.3333
% notre covariance
n=length(x)
moy_x = mean(x)
moy_y = mean(y)
my_cov = sum( (x-moy_x).*(y-moy_y))/n % ans = 1.3333
```

La covariance joue un rôle analogue à la variance dans le cas de deux caractères : elle mesure la dispersion conjointe des deux caractères. La corrélation joue un rôle analogue à l'écart type.

En supposant $V(\mathbf{x}) > 0$ et $V(\mathbf{y}) > 0$, c'est-à-dire que $n \geq 2$ et les x_k (resp. les y_k) ne sont pas tous égaux, on peut définir le **coefficient de corrélation linéaire (de Bravais-Pearson)** :

$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}}.$$

On a

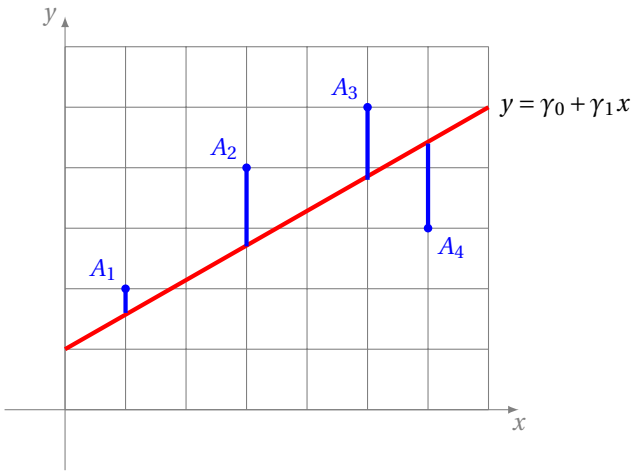
- $r(\lambda\mathbf{x}, \lambda\mathbf{y}) = r(\mathbf{x}, \mathbf{y})$ pour tout $\lambda \in \mathbb{R}^*$,
- $r(\mathbf{x}, \mathbf{y}) \in [-1; 1]$.

1.5 Régression linéaire revisitée

L'ANALYSE DE RÉGRESSION donne des outils de prédiction du comportement d'un caractère si on connaît la valeur d'un autre caractère. L'ANALYSE DE CORRÉLATION mesure la force de la relation linéaire entre les deux caractères.
Considérons une série statistique bivariée $(x_k, y_k)_{k\in[\![1;n]\!]}$. On peut associer à chaque donnée (x_k, y_k) un point du plan et on peut représenter un échantillon de n données comme un nuage de n points. Si le nuage a une forme allongée, on peut essayer de dessiner une droite passant au milieu de ces points. Cette droite, appelée droite de régression linéaire, est un moyen de représenter la dépendance linéaire des deux caractères. La méthode des moindre carrés permet de déterminer la "meilleur" droite passante par le nuage de points constitué par une série statistique double.

1.5.1 Régression linéaire et moindres carrés

On considère un ensemble de N points $A_i = (x_i, y_i)$, $i = 1, \dots, N$. L'objectif est de trouver l'équation $y = \gamma_0 + \gamma_1 x$ de la droite qui approche au mieux tous ces points. Précisons ce que veut dire "approcher au mieux" : il s'agit de minimiser la somme des carrés des distances verticales entre les points et la droite.



La formule qui donne l'erreur est :

$$E(\gamma_0, \gamma_1) = \sum_{i=1}^{N} \big(y_i - (\gamma_0 + \gamma_1 x)\big)^2,$$

autrement dit

$$E(\gamma_0, \gamma_1) = \big(y_1 - (\gamma_0 + \gamma_1 x_1)\big)^2 + \cdots + \big(y_N - (\gamma_0 + \gamma_1 x_N)\big)^2.$$

Remarquons que l'on a toujours $E(\gamma_0, \gamma_1) \geq 0$. Si par exemple tous les points sont alignés, alors on peut trouver a et b tels que $E(\gamma_0, \gamma_1) = 0$. Quand ce n'est pas le cas, on cherche γ_0 et γ_1 qui rendent $E(\gamma_0, \gamma_1)$ le plus petit possible. Il s'agit donc bien ici de minimiser une fonction de deux variables (les variables sont γ_0 et γ_1). Pour cela nous aurons besoin de calculer son gradient :

$$\nabla E(\gamma_0, \gamma_1) = \left(\frac{\partial E}{\partial \gamma_0}(\gamma_0, \gamma_1)\right), \frac{\partial E}{\partial \gamma_1}(\gamma_0, \gamma_1), = \left(\sum_{i=1}^{N} -2\big(y_i - (\gamma_0 + \gamma_1 x_i)\big), \sum_{i=1}^{N} -2x_i\big(y_i - (\gamma_0 + \gamma_1 x_i)\big)\right).$$

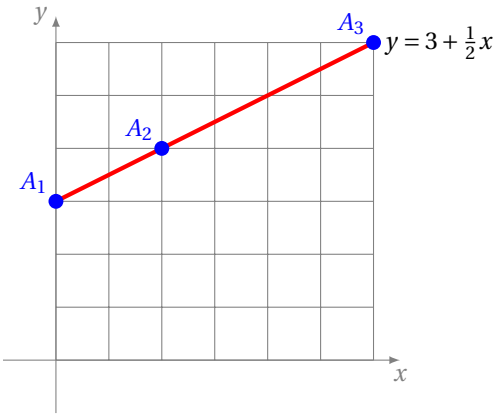
EXEMPLE
Prenons d'abord l'exemple des trois points $A_1 = (0,3)$, $A_2 = (2,4)$ et $A_3 = (6,6)$. La fonction $E(\gamma_0, \gamma_1)$ s'écrit :

$$E(\gamma_0, \gamma_1) = (3 - \gamma_0)^2 + (4 - (\gamma_0 + 2\gamma_1))^2 + (6 - (\gamma_0 + 6\gamma_1))^2 = 40\gamma_1^2 + 16\gamma_1\gamma_0 - 88\gamma_1 + 3\gamma_0^2 - 26\gamma_0 + 61.$$

Ainsi

$$\nabla E(\gamma_0, \gamma_1) = \begin{pmatrix} 16\gamma_1 + 6\gamma_0 - 26 \\ 80\gamma_1 + 16\gamma_0 - 88 \end{pmatrix}$$

et $\nabla E(\gamma_0, \gamma_1) = \mathbf{0}$ ssi $\gamma_1 = \frac{1}{2}$ et $\gamma_0 = 3$. De plus, $E(3, \frac{1}{2}) = 0$ (les points sont alignés).



EXEMPLE
À partir des données des 5 points suivants, quelle ordonnée peut-on extrapoler pour le point d'abscisse $x = 6$?

$$A_1 = (4,1), \quad A_2 = (7,3), \quad A_3 = (8,3), \quad A_4 = (10,6), \quad A_5 = (12,7).$$

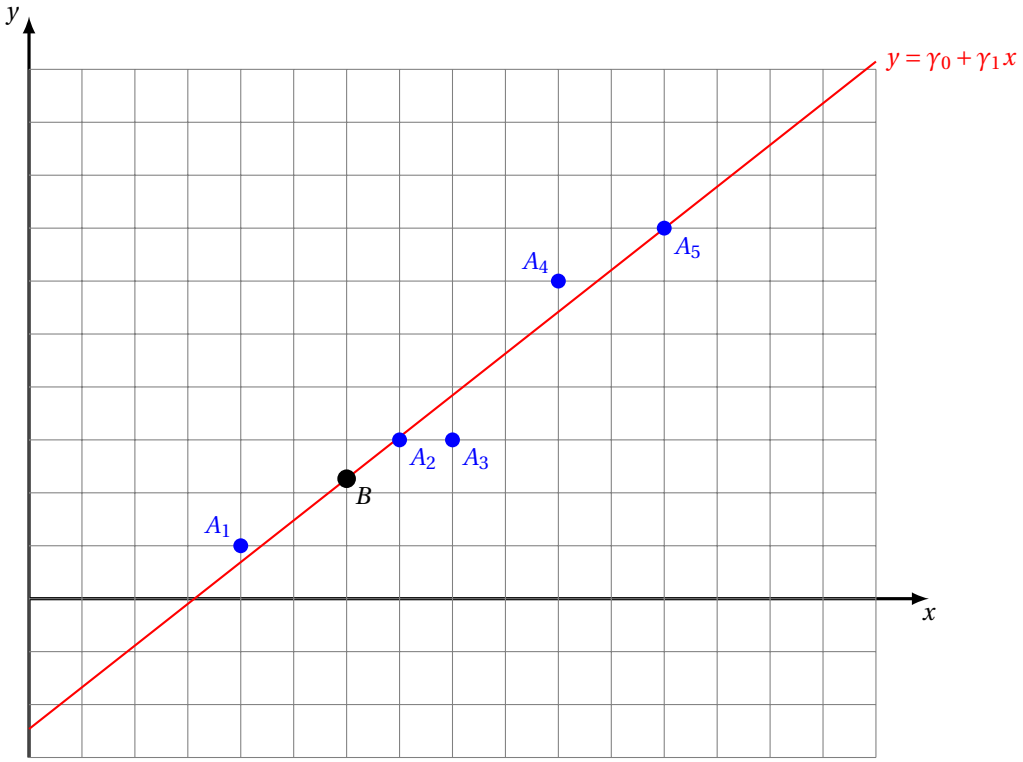
Ces 5 points sont à peu près alignés. On calcule la meilleure droite de régression linéaire en minimisant la fonction $E(\gamma_0, \gamma_1)$:

$$\begin{aligned} E(\gamma_0,\gamma_1) &= (-12\gamma_1 - \gamma_0 + 7)^2 + (-10\gamma_1 - \gamma_0 + 6)^2 + (-8\gamma_1 - \gamma_0 + 3)^2 + (-7\gamma_1 - \gamma_0 + 3)^2 + (-4\gamma_1 - \gamma_0 + 1)^2 \\ &= 373\gamma_1^2 + 82\gamma_1\gamma_0 - 386\gamma_1 + 5\gamma_0^2 - 40\gamma_0 + 104746\gamma_1 + 82\gamma_0 - 386. \end{aligned}$$

Ainsi

$$\nabla E(\gamma_0,\gamma_1) = \begin{pmatrix} 82\gamma_1 + 10\gamma_0 - 40 \\ 746\gamma_1 + 82\gamma_0 - 386 \end{pmatrix}$$

et $\nabla E(\gamma_0, \gamma_1) = \mathbf{0}$ ssi $\gamma_1 = \frac{145}{184} \approx 0.788$ et $\gamma_0 = -\frac{453}{184} \approx -2.462$. De plus, $E(-\frac{453}{184}, \frac{145}{184}) = 211/184 > 0$ (les points ne sont pas alignés).
Par conséquent, selon notre modèle linéaire, pour $x = 6$, on doit avoir $y = \gamma_0 + 6\gamma_1 = \frac{417}{184} \approx 2.27$ (le point B de la figure ci-dessus).



1.5.2 Droite de régression de y par rapport à x

On cherche à déterminer la droite d'équation $y = \gamma_0 + \gamma_1 x$ minimisant l'erreur quadratique $\mathscr{E} : \mathbb{R}^2 \to \mathbb{R}_+$ définie par

$$\mathscr{E}(\gamma_0, \gamma_1) = \sum_{k=1}^{n} \big(y_k - (\gamma_0 + \gamma_1 x_k)\big)^2$$

qui est la somme des distances au carré entre les points (x_k, y_k) et les points $(x_k, \gamma_0 + \gamma_1 x_k)$ de même abscisse situés sur la droite $y = \gamma_0 + \gamma_1 x$. Au chapitre **??** on a montré que γ_0 et γ_1 sont solution du système linéaire [1]

$$\begin{bmatrix} 1 & \frac{1}{n}\sum_{k=1}^n x_k \\ \frac{1}{n}\sum_{k=1}^n x_k & \frac{1}{n}\sum_{k=1}^n x_k^2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n y_k \\ \frac{1}{n}\sum_{k=1}^n x_k y_k \end{bmatrix}$$

autrement dit, avec les notations introduites dans ce chapitre,

$$\begin{bmatrix} 1 & \bar{\mathbf{x}} \\ \bar{\mathbf{x}} & V(\mathbf{x}) + (\bar{\mathbf{x}})^2 \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{y}} \\ C(\mathbf{x}, \mathbf{y}) + \bar{\mathbf{x}}\bar{\mathbf{y}} \end{bmatrix}.$$

En résolvant ce système on trouve

$$\gamma_1 = \frac{C(\mathbf{x}, \mathbf{y})}{V(\mathbf{x})}, \qquad \text{coefficient directeur (pente)},$$
$$\gamma_0 = \bar{\mathbf{y}} - \gamma_1 \bar{\mathbf{x}}, \qquad \text{ordonnée à l'origine},$$

autrement dit $y = \gamma_1(x - \bar{\mathbf{x}}) + \bar{\mathbf{y}}$ (la droite passe par le point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$).
D'un point de vue computationnel, cette écriture est susceptible de générer des erreurs de *roundoff* (les deux termes au numérateur ainsi qu'au dénominateur sont presque égaux, *i.e.* $C(\mathbf{x}, \mathbf{y})$ et $V(\mathbf{x})$ sont proches de zéro). Il est alors plus stable de calculer γ_1 comme suit (ce qui est équivalent) :

$$\gamma_1 = \frac{\sum_{k=0}^n \big(y_k(x_k - \bar{\mathbf{x}})\big)}{\sum_{k=0}^n \big(x_k(x_k - \bar{\mathbf{x}})\big)}.$$

1.5.3 Droite de régression de x par rapport à y

En échangeant les rôles de $\mathbf{x}$ et $\mathbf{y}$ on obtient la régression linéaire de $\mathbf{x}$ par rapport à $\mathbf{y}$. En générale les deux droites de régression sont distinctes.
En effet, dans le premier cas on minimise la somme des distances "verticales" (*i.e.* à x_i fixé), dans le deuxième cas il s'agit des distances "horizontale" (*i.e.* à y_i fixé) et en général ces deux quantités sont différentes.
Le produit des pentes de ces deux droites est égal à r^2 et les deux pentes sont égales si et seulement si $r = \pm 1$. Dans ce cas les deux droites coïncident et les points sont alignés.

1.5.4 Interprétation du coefficient de corrélation linéaire r

Il est toujours possible de tracer la droite des moindres carrés quelle que soit la forme du nuage. L'approximation du nuage par cette droite est-elle légitime ? Quel sens, quelle signification donner à cette droite ?
Dans un ajustement linéaire de $\mathbf{y}$ par rapport à $\mathbf{x}$ on appelle $\mathbf{x}$ la variable explicative (ou le "prédicteur") et $\mathbf{y}$ la variable expliquée (ou "à expliquer"). Le but d'un ajustement linéaire est d'expliquer une partie de la variation de $\mathbf{y}$ du fait de sa dépendance linéaire à $\mathbf{x}$.
Nous allons voir que le coefficient de corrélation r peut être utilisé pour mesurer la qualité d'une approximation de $\mathbf{y}$ par une fonction linéaire en x. Lorsque $r(\mathbf{x}, \mathbf{y})$ est en valeur absolue proche de 1 (en pratique strictement supérieur à 0.7), la droite de régression linéaire est une bonne approximation du nuage de point.

Notons $\hat{y}_k = \gamma_0 + \gamma_1 x_k$ pour $k = 1, \ldots, n$ la valeur estimée (ou prédite ou ajustée) de y_k par la régression linéaire lorsque $x = x_k$ et $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. Il semble naturel de dire que remplacer le nuage par la droite trouvée est d'autant plus légitime que la dispersion du nuage de points par rapport à la droite des moindres carrés est petite. Autrement dit, on calcul l'erreur quadratique en son minimum (γ_0, γ_1) : l'approximation est légitime plus l'erreur quadratique $\mathscr{E}(\gamma_0, \gamma_1)$ est faible.

Soit γ_0 et γ_1 les valeurs qui minimisent l'erreur quadratique, alors

$$n\mathscr{E}(\gamma_0, \gamma_1) = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

1. NB : ici les indices commencent à 1 et on a tout divisé par n.

$$
\begin{aligned}
&= n\sum_{k=1}^{n}(y_k-\gamma_0-\gamma_1 x_k)^2\\
&= n\sum_{k=1}^{n}(y_k-(\bar{\mathbf{y}}-\gamma_1\bar{\mathbf{x}})-\gamma_1 x_k)^2\\
&= n\sum_{k=1}^{n}((y_k-\bar{\mathbf{y}})-\gamma_1(x_k-\bar{\mathbf{x}}))^2\\
&= n\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})^2+n\gamma_1^2\sum_{k=1}^{n}n(x_k-\bar{\mathbf{x}})^2-2n\gamma_1\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})(x_k-\bar{\mathbf{x}})\\
&= V(\mathbf{y})+\gamma_1^2 V(\mathbf{x})-2\gamma_1 C(\mathbf{x},\mathbf{y})\\
&= V(\mathbf{y})+\frac{C^2(\mathbf{x},\mathbf{y})}{V^2(\mathbf{x})}V(\mathbf{x})-2\frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})}C(\mathbf{x},\mathbf{y})\\
&= V(\mathbf{y})-\frac{C^2(\mathbf{x},\mathbf{y})}{V(\mathbf{x})}=V(\mathbf{y})\left(1-r^2(\mathbf{x},\mathbf{y})\right).
\end{aligned}
$$

Qualitativement, plus cette erreur est grande et moins bon est l'ajustement linéaire obtenu.
La quantité

$$SC_{\text{rés}}\overset{\text{def}}{=}\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

est appelée **somme des carrés résiduelle** et est donc égale à

$$SC_{\text{rés}}=n\mathscr{E}(\gamma_0,\gamma_1)=V(\mathbf{y})\left(1-r^2(\mathbf{x},\mathbf{y})\right).$$

Elle est d'autant plus faible que r^2 est proche de 1. On peut alors interpréter l'erreur quadratique comme une mesure de la part de la variance de $\mathbf{y}$ qui ne peut pas être expliquée et prédite par une fonction linéaire en $\mathbf{x}$.
La variation totale

$$SC_{\text{tot}}\overset{\text{def}}{=}\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})^2$$

est appelée **somme des carrés totale** de $\mathbf{y}$ et est égale à

$$SC_{\text{tot}}=nV(\mathbf{y}).$$

On a donc

$$1-r^2(\mathbf{x},\mathbf{y})=\frac{SC_{\text{rés}}}{SC_{\text{tot}}},$$

i.e. la quantité $(1-r^2(\mathbf{x},\mathbf{y}))$ est égale à la proportion de variation de $\mathbf{y}$ non expliquée par la droite des moindres carrés.
La décomposition de la variation totale de $\mathbf{y}$ permet une autre interprétation de r^2 :

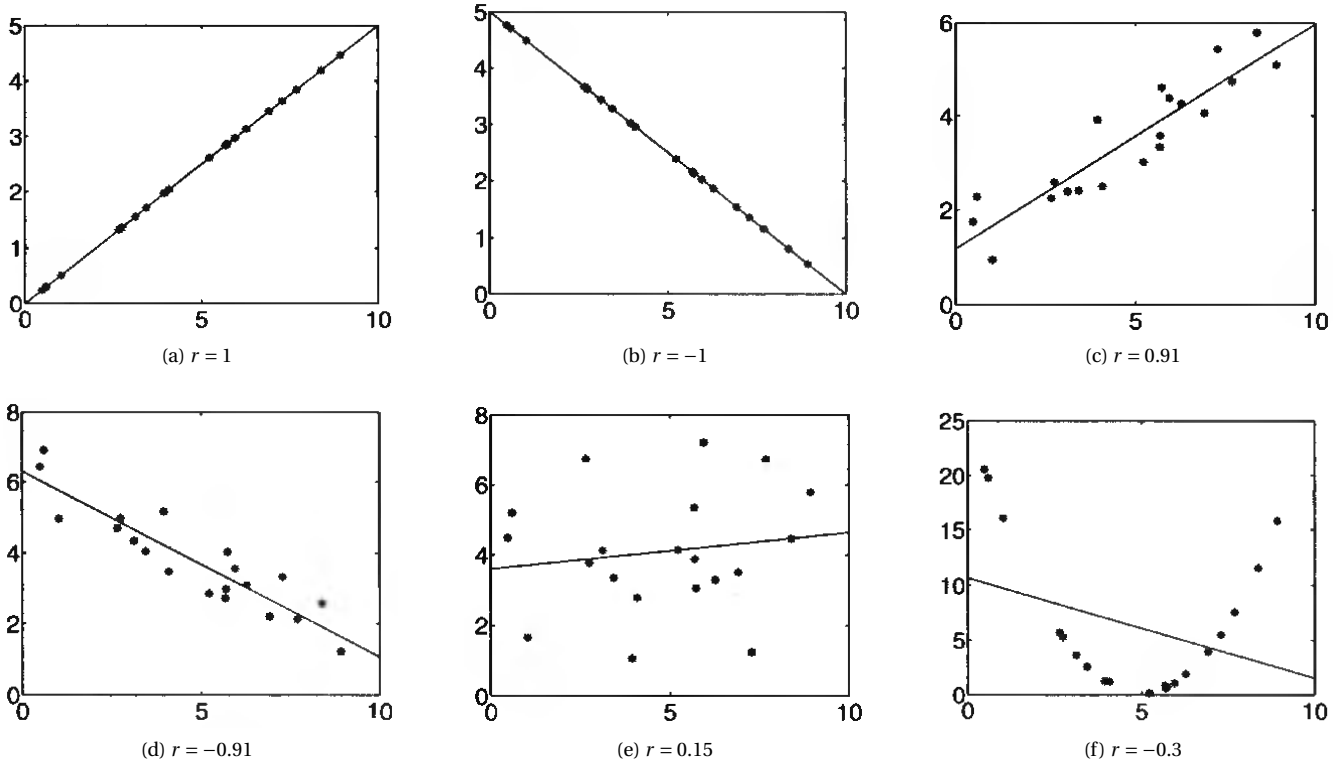
$$
\begin{aligned}
SC_{\text{tot}} &= \sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})^2\\
&= \sum_{k=1}^{n}(y_k-\hat{y}_k+\hat{y}_k-\bar{\mathbf{y}})^2\\
&= \sum_{k=1}^{n}(y_k-\hat{y}_k)^2+\sum_{k=1}^{n}(\hat{y}_k-\bar{\mathbf{y}})^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{\mathbf{y}})\\
&= SC_{\text{rés}}+\sum_{k=1}^{n}(\hat{y}_k-\bar{\mathbf{y}})^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{\mathbf{y}}).
\end{aligned}
$$

Montrons que le dernier terme est nul :

$$
\begin{aligned}
\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{\mathbf{y}}) &= \gamma_1\left(\sum_{k=1}^{n}(y_k-\bar{\mathbf{y}})(x_k-\bar{\mathbf{x}})-\gamma_1\sum_{k=1}^{n}(x_k-\bar{\mathbf{x}})^2\right)\\
&= \gamma_1\left(C(\mathbf{x},\mathbf{y})-\gamma_1 V(\mathbf{x})\right)=0.
\end{aligned}
$$

On appelle **variation expliquée** par la régression la quantité

$$SC_{\text{expl}}=\sum_{k=1}^{n}(\hat{y}_k-\bar{\mathbf{y}})^2=nV(\hat{\mathbf{y}}).$$



(a) $r = 1$ (b) $r = -1$ (c) $r = 0.91$

(d) $r = -0.91$ (e) $r = 0.15$ (f) $r = -0.3$

FIGURE 1.3 – Qualité des ajustements linéaires en fonction du coefficient de corrélation. Ce coefficient reflète la non-linéarité et la direction d'une relation linéaire mais pas la pente de cette relation ni de nombreux aspects des relations non linéaires (dernière figure).

et on a l'égalité

$$SC_{\text{tot}} = SC_{\text{rés}} + SC_{\text{expl}}.$$

On a donc

$$r^2(\mathbf{x},\mathbf{y}) = \frac{SC_{\text{expl}}}{SC_{\text{tot}}},$$

i.e. $r^2(\mathbf{x},\mathbf{y})$ est égale à la proportion de variation de $\mathbf{y}$ expliquée par la droite des moindres carrés.

Le coefficient de corrélation r mesure la force et la direction de la relation entre $\mathbf{x}$ et $\mathbf{y}$. Deux cas extrêmes peuvent être facilement analysés :

- si $r(\mathbf{x},\mathbf{y}) = \pm 1$, alors il existe un $\lambda_0 \in \mathbb{R}^*$ tel que $y_k - \bar{\mathbf{y}} = \lambda_0(x_k - \bar{\mathbf{x}})$ pour tout $k \in [\![1;n]\!]$. Cela montre que $\mathbf{x}$ et $\mathbf{y}$ sont parfaitement corrélés ;
- si $r(\mathbf{x},\mathbf{y}) = 0$, alors la meilleur droite d'ajustement linéaire est la droite horizontale d'équation $y = \bar{\mathbf{y}}$ ce qui tend à montrer que les deux caractères ne sont pas corrélés.

La figure 1.3 donne plusieurs exemples pour différentes valeurs du coefficient de corrélation. Une valeur de $r(\mathbf{x},\mathbf{y})$ proche de 1 indique que les caractères sont positivement corrélés, et la meilleure droite d'ajustement linéaire obtenue par la méthode des moindres carrés a une pente positive. Une valeur de $r(\mathbf{x},\mathbf{y})$ proche de -1 indique que les caractères sont négativement corrélés, et la meilleure droite d'ajustement linéaire a une pente négative.

Noter que le coefficient de corrélation mesure seulement la qualité d'une relation linéaire : les caractères peuvent être corrélés mais pas linéairement, dans ce cas r sera petit et il faudrait généraliser ces notions aux cas des ajustements polynomiales.

Remarque (r : diviser par n ou $n-1$?)

Dans la définition de r, les dénominateurs utilisés pour la covariance $C(\mathbf{x},\mathbf{y})$ et pour les variances $V(\mathbf{x})$ et $V(\mathbf{y})$ sont n.

Si l'on tente d'estimer la corrélation de la population à partir d'un échantillon il faudra utiliser l'estimation de la covariance, toujours notée $C(\mathbf{x},\mathbf{y})$, ainsi que les estimations des variances $E(V(\mathbf{x}))$ et $E(V(\mathbf{y}))$, cela revient à diviser par $(n-1)$. Ce rapport

donne la même valeur que r :

$$\frac{E(C(\mathbf{x},\mathbf{y}))}{\sqrt{E(V(\mathbf{x}))E(V(\mathbf{y}))}} = \frac{\frac{n}{n-1}C(\mathbf{x},\mathbf{y})}{\sqrt{\frac{n}{n-1}V(\mathbf{x})\frac{n}{n-1}V(\mathbf{y})}} = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = r(\mathbf{x},\mathbf{y}).$$

1.6 Corrélation et mises en garde

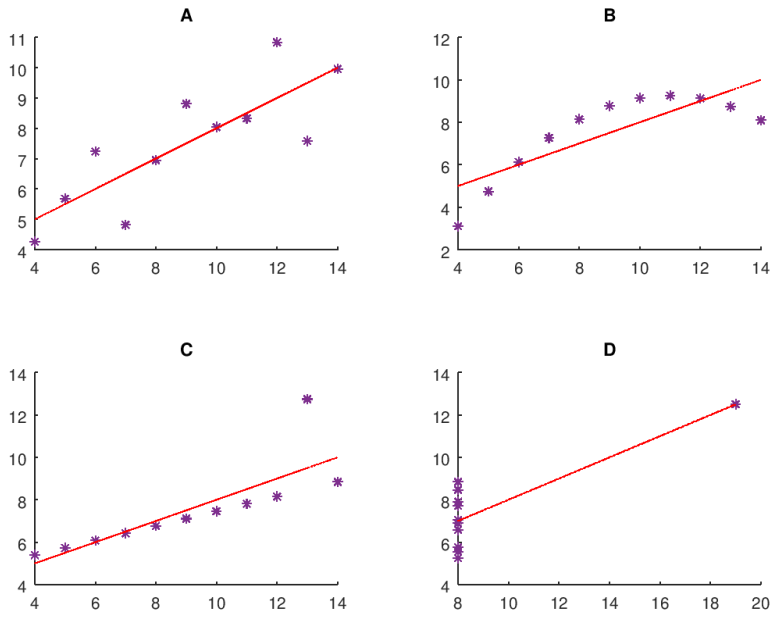
1.6.1 Le coefficient r et la qualité de l'ajustement linéaire

Comment juger la qualité de l'ajustement linéaire? Il est clair que si le coefficient r est voisin de 0, il faut rejeter l'ajustement linéaire, mais pour quelles valeurs de r, le considère-t-on de bonne qualité? C'est une question importante et beaucoup d'exemples montrent qu'on ne peut pas établir de règles de décision à partir du seul examen de la valeur de r.
Les exemples suivants montrent que le calcul du coefficient de corrélation linéaire doit toujours être complété par un examen graphique. Pour d'autres exemples voir par exemple https://www.autodesk.com/research/publications/same-stats-different-graphs

EXEMPLE
Considérons les quatre séries de 11 observations simultanées de deux variables **x** et **y** suivantes :

Série A		Série B		Série C		Série D	
x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	8.00	5.56
12.00	10.84	12.00	9.13	12.00	8.15	19.00	12.50
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

On obtient grosso modo la même valeur du coefficient de corrélation linéaire ($r \approx 0.816$) et la même droite des moindres carrés $y \approx 3 + 0.5x$, mais l'examen graphique montre que l'ajustement linéaire n'est adapté qu'au premier cas.



EXEMPLE

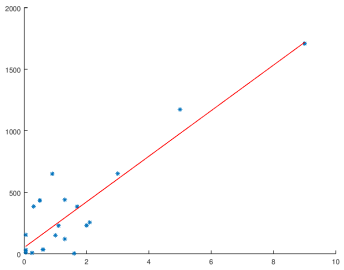
On se propose de calculer l'ajustement linéaire de la série de la composition minérale en fluorures et sodium (mg/l) de 21 eaux minérales gazeuses : [2]

Eau minérale	$\mathbf{x}$ = Fluorures	$\mathbf{y}$ = Sodium
Arcens	1.3	439
Arvie	0.9	650
Badoit	1	150
Beckerich	0.6	34
Châteauneuf	3	651
Eau de Perrier	0.05	11.5
Faustine	2	230
La Salvetat	0.25	7
Perrier	0.05	11.5
Puits St-Georges	0.5	434
Pyrénées	0.05	31
Quézac	2.1	255
San Pellegrino	0.6	35
St-Diéry	0.3	385
St-Jean	1.1	228
St-Pierre	1.7	383
St-Yorre	9	1708
Vernet	1.3	120
Vernière	0.05	154
Vichy-Célestins	5	1172
Wattwiller	1.6	3

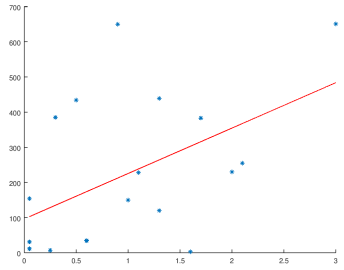
Calculons tout d'abord la moyenne et l'écart type :

$$\bar{\mathbf{x}} = 1.55, \qquad \bar{\mathbf{y}} = 338,$$
$$\sigma(\mathbf{x}) = 2.03, \qquad \sigma(\mathbf{x}) = 417.$$

Le coefficient de corrélation linéaire entre les deux composants minéraux est égal à 0.90. Cette valeur assez proche de 1 peut conduire à considérer que la droite des moindres carrés permet d'évaluer approximativement la teneur y en sodium en fonction de la teneur x en fluorures :



Cependant la représentation graphique du nuage des 21 points montre deux points caractérisés par une minéralité particulièrement élevée : «Vichy-Célestins» et «Saint-Yorre». Ces deux eaux minérales ont respectivement des valeurs « éloignée » et « extrême » pour les deux composants minéraux.En supprimant ces deux points et en réalisant l'ajustement sur les 19 autres points, on obtient :



2. Données extraites du journal "Que Choisir?", n° 422 bis, 2005

La moyenne et l'écart type sont maintenant

$$\bar{\mathbf{x}} = 0.97, \qquad \bar{\mathbf{y}} = 222,$$
$$\sigma(\mathbf{x}) = 0.81, \qquad \sigma(\mathbf{x}) = 208$$

et le coefficient r est passé de 0.9 à 0.5. Il faut aussi remarquer que les coefficients de la droite des moindres carrés sont passés respectivement de 185 à 129 et de 51 à 96.15.
Quel crédit apporter à un ajustement pour lequel deux points ont une telle influence ? On est donc obligé d'abandonner l'idée d'une relation linéaire entre les deux composants minéraux.

Tous ces résultats montrent qu'il ne faut jamais conclure sur la dépendance entre deux variables quantitatives au seul examen de la valeur du coefficient de corrélation linéaire.

De plus, lorsqu'une liaison linéaire entre deux variables a été mise en évidence par l'étude d'une série de n observations sur ce couple, il faut bien se garder de conclure à une relation de cause à effet entre ces variables sans en avoir examiné attentivement la signification, comme on va voir à la prochaine section.
L'examen graphique, ainsi que celui de la signification des variables, sont des compléments indispensables à l'information donnée par la valeur du coefficient de corrélation linéaire.

1.6.2 Distinguer causalité et corrélation

En statistiques, deux variables (choses que l'on mesure) sont corrélées positivement si elles évoluent de la même façon (augmentent en même temps, diminuent en même temps). Elles sont corrélées négativement si elles évoluent en sens inverse.
On établie un lien de causalité entre deux variables lorsqu'il y a un lien de cause à effet entre les deux, lorsque l'une est conséquence de l'autre.
L'**effet cigogne**[3] est une erreur qui consiste à confondre corrélation et causalité : «Deux variables évoluent de la même façon, l'une est donc forcément la cause de l'autre».

> «L'Alsace est la région de France où l'on observe le plus de cigognes. C'est également la région de France où il y a le plus de naissances. C'est donc la preuve que les cigognes apportent les bébés.»

Erreur si proche de l'effet cigogne qu'on les confond souvent, il s'agit ici de confondre succession et causalité[4] : «Deux évènements se suivent dans le temps, le premier est donc forcément la cause du second.»

EXEMPLE
Voici quelques exemples de ces deux confusions.

- Thomas met son caleçon rayé, puis il va au casino et gagne le gros lot. Il en conclue que son caleçon lui a porté chance.
- Plus les éoliennes tournent vite, plus y il a du vent : ce sont donc les éoliennes qui créent le vent !
- On constate que les pays où l'on mange le plus de viande sont les pays où l'on vit le plus longtemps. Doit-je changer mon régime alimentaire ? (On constate en réalité que ces pays sont également les plus riches, donc ceux où les habitants peuvent à la fois acheter plus de viande et avoir accès à de meilleurs soins)
- On constate que depuis que le parti de M. X est au pouvoir, le chômage diminue. Dois-je voter pour lui aux prochaines élections ? (Le chômage est lié à un grand nombre de facteurs très complexes, une simple corrélation est donc insuffisante pour démontrer que les actions de ce parti sont la cause de cette diminution. Il y a probablement un grand nombre de causes.)
- Je traînais un gros rhume depuis 3 jours, j'ai pris une tisane de camomille et le lendemain, j'allais mieux. La camomille m'a-t-elle guérie ? Ou bien est-ce j'aurais guéri de la même façon sans prendre de tisane, parce qu'un rhume se soigne généralement tout seul en 3 jours ?

Bien entendu, un corrélation peut donner des indices, interroger. Mais il ne s'agit en aucun cas d'un fait suffisant pour démontrer un lien de causes à effets. Pourtant, le raccourcit est rapide, instinctif, très largement utilisé dans les médias, et parfois très dangereux.
Une corrélation et une causalité sont deux objets distincts. Deux événements peuvent être corrélés sans pour autant avoir des rapports de cause à effet car d'autres variables pourraient être la cause des variations de **x** et de **y**.
Considérons par exemple l'affirmation suivante due à Coluche :

> «Quand on est malade, il ne faut surtout pas aller à l'hôpital : la probabilité de mourir dans un lit d'hôpital est 10 fois plus grande que dans son lit à la maison».

3. ou *Cum hoc, ergo propter hoc* : avec cela, donc à cause de cela.
4. ou *Post hoc, ergo propter hoc* : après cela, donc à cause de cela

Or, on ne meurt pas plus parce qu'on est dans un lit d'hôpital, mais on y est parce qu'on est malade, et quand on est malade la probabilité de mourir est plus grande.

Un autre exemple : une étude anglaise a prouvé que les gens habitant près de pylônes à haute tension étaient significativement plus souvent malades que le reste de la population. Est-ce la faute du courant électrique ? Ce n'est pas évident parce qu'une autre étude a révélé que les habitants sous les pylônes étaient en moyenne plus pauvres et on sait la corrélation (causalité ?) santé-pauvreté. À elle seule, cette étude ne permet pas de conclure.

Il en va ainsi des corrélations délinquance et origine ethnique : même à supposer qu'elles soient vraies, elles ne démontrent pas le rapport de cause à effet ; il peut se faire que la pauvreté, voire la détresse, soient liées à des discriminations ethniques, c'est alors cette misère qui est une cause possible de délinquance.

Démontrer une théorie avec seulement des statistiques peut être trompeur. Souvent la théorie préexiste et les chiffres sont ensuite utilisés pour la conforter «scientifiquement».

La corrélation relie les données et c'est ce que les big data brassent à très grosse échelle aujourd'hui. Ils accumulent une somme considérable de données et ils croisent tout ça en fonction de ce que l'on veut faire dire. Cependant, pour déterminer la nature du lien de causalité entre plusieurs éléments, c'est plus complexe. La théorie doit avoir un pouvoir explicatif, ne serait-ce que pour savoir dans quel sens lire les corrélations si jamais un lien de causalité existe. Il est par exemple maintenant bien établi qu'historiquement les variations de température sont étroitement liées aux variations de concentration de gaz carbonique dans l'atmosphère. Mais c'est la théorie qui permet de dire si c'est le réchauffement qui crée l'excès de gaz carbonique, ou l'inverse.

1.7 Exercices

Exercice 1.1 (Série univariée)
Une classe a été divisée en deux groupes de TP : le groupe TP_1 de $n_1 = 10$ étudiants et le groupe TP_2 de $n_2 = 4$ étudiants. Lors d'un contrôle noté sur 5, les étudiants du groupe TP_1 ont reçu les notes 4, 1, 3, 3, 4, 2, 3, 5, 3, 4 tandis que ceux du groupe TP_2 ont reçu les notes 4, 4, 4 et 5.
Pour chaque groupe κ, calculer la moyenne, le mode et la médiane des notes.
Calculer ensuite la moyenne, le mode et la médiane des notes de la classe.

Correction
Pour le groupe TP_1 on note $\mathbf{u} = (1,2,3,3,3,3,4,4,4,5)$ (dans l'ordre croissante) ce qui donne le tableau des fréquences

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	1	1/10
2	1	1/10
3	4	4/10
4	3	3/10
5	1	1/10
	$\sum = 10$	$\sum = 1$

Le mode est 3 (c'est la classe la plus importante). La moyenne vaut

$$\bar{\mathbf{u}} = \frac{1}{10}(1+2+3+3+3+3+4+4+4+5) = 3.2$$

soit encore, à partir du tableau,

$$\bar{\mathbf{u}} = \frac{1}{10}(1\times 1+2\times 1+3\times 4+4\times 3+5\times 1) = 3.2$$

Comme on a un nombre pair d'éléments (10), la médiane vaut

$$M(\mathbf{u}) = \frac{u_5+u_6}{2} = 3.$$

Pour le groupe TP_2 on note $\mathbf{v} = (4,4,4,5)$ (dans l'ordre croissante) ce qui donne le tableau des fréquences

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	0	0/10
2	0	0/10
3	0	0/10
4	3	3/10
5	1	1/10
	$\sum = 4$	$\sum = 1$

Le mode est 4 (c'est la classe la plus importante). La moyenne vaut

$$\bar{\mathbf{v}} = \frac{1}{4}(4+4+4+5) = 4.25$$

soit encore, à partir du tableau,

$$\bar{\mathbf{v}} = \frac{1}{4}(1\times 0 + 2\times 0 + 3\times 0 + 4\times 3 + 5\times 1) = 4.25$$

Comme on a un nombre pair d'éléments (4), la médiane vaut

$$M(\mathbf{v}) = \frac{v_2 + v_3}{2} = 4.$$

Pour la classe fusion des deux groupes de TP, on note $\mathbf{x} = (1,2,3,3,3,3,4,4,4,4,4,4,5,5)$ (dans l'ordre croissante) ce qui donne le tableau des fréquences

Note	Effectif (Nombre d'étudiants)	Fréquence (Proportion d'étudiants)
1	1	1/14
2	1	1/14
3	4	4/14
4	6	6/14
5	2	2/14
	$\sum = 14$	$\sum = 1$

Le mode est 4 (c'est la classe la plus importante). La moyenne vaut

$$\bar{\mathbf{x}} = \frac{1}{14}(1\times 1 + 2\times 1 + 3\times 4 + 4\times 6 + 5\times 2) = \frac{49}{14} = 3.5$$

soit encore, d'après la propriété sur la fusion de données,

$$\bar{\mathbf{x}} = \frac{n_1\bar{\mathbf{u}} + n_2\bar{\mathbf{v}}}{n_1+n_2} = \frac{10\times 3.2 + 4\times 4.25}{10+4} = \frac{49}{14} = 3.5$$

Comme on a un nombre pair d'éléments (14), la médiane vaut

$$M(\mathbf{x}) = \frac{x_7 + x_8}{2} = 4.$$

Exercice 1.2 (Covariance)

Calculer la covariance dans les cas suivants :

1. $\{(1,1),(-1,-1)\}$
2. $\{(-1,1),(1,-1)\}$
3. $\{(1,1),(-1,-1),(-1,1),(1,-1)\}$

Correction

1. On a $\mathbf{x} = (1,-1)$ et $\mathbf{y} = (1,-1)$ donc $\bar{\mathbf{x}} = \bar{\mathbf{y}} = 0$ et $\mathbb{C}(\mathbf{x},\mathbf{y}) = \frac{1\times 1 + (-1)\times(-1)}{2} - 0 = 1$: les points sont alignés et la pente est positive.
2. On a $\mathbf{x} = (-1,1)$ et $\mathbf{y} = (1,-1)$ donc $\bar{\mathbf{x}} = \bar{\mathbf{y}} = 0$ et $\mathbb{C}(\mathbf{x},\mathbf{y}) = \frac{(-1)\times 1 + 1\times(-1)}{2} - 0 = -1$: les points sont alignés et la pente est négative.

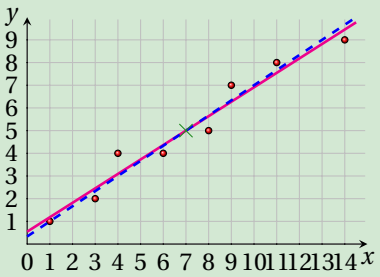
3. On a $\mathbf{x}=(1,-1,-1,1)$ et $\mathbf{y}=(1,-1,1,-1)$ donc $\bar{\mathbf{x}}=\bar{\mathbf{y}}=0$ et $\mathbb{C}(\mathbf{x},\mathbf{y})=\frac{1\times1+(-1)\times(-1)+(-1)\times1+1\times(-1)}{4}-0=0$: il n'y a pas de corrélation.

Exercice 1.3 (Régression linéaire)

Calculer les droites de meilleur approximation de l'ensemble de points suivant :

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

ainsi que leurs coefficients de corrélation.



Correction

Nous avons une série statistique double avec une population d'effectif $n=8$.
Pour calculer la droite de régression de $\mathbf{y}$ par rapport à $\mathbf{x}$ on calcule les quantités suivantes :

$$\bar{\mathbf{x}}=\frac{1}{n}\sum_{k=1}^{n}x_k=\frac{56}{8}=7$$
$$\bar{\mathbf{y}}=\frac{1}{n}\sum_{k=1}^{n}y_k=\frac{40}{8}=5$$
$$V(\mathbf{x})=\frac{1}{n}\sum_{k=1}^{n}x_k^2-\bar{\mathbf{x}}^2=\frac{33}{2}$$
$$C(\mathbf{x},\mathbf{y})=\frac{1}{n}\sum_{k=1}^{n}x_k y_k-\bar{\mathbf{x}}\bar{\mathbf{y}}=\frac{21}{2}$$
$$\gamma_1=\frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})}=\frac{7}{11}$$
$$\gamma_0=\bar{\mathbf{y}}-\gamma_1\bar{\mathbf{x}}=\frac{6}{11}$$
$$V(\mathbf{y})=\frac{1}{n}\sum_{k=1}^{n}y_k^2-\bar{\mathbf{y}}^2=7$$
$$r(\mathbf{x},\mathbf{y})=\frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}}=\sqrt{\frac{21}{22}}>0.97$$

La droite cherchée a donc pour équation $y=\gamma_0+\gamma_1 x=\frac{6}{11}+\frac{7}{11}x$ avec une forte corrélation (mais cela ne dit rien sur la causalité entre les deux quantités!).

```
xx=[1,3,4,6,8,9,11,14]
yy=[1,2,4,4,5,7,8,9]

n = length(xx)
moy_x = mean(xx) %sum(xx)/n
moy_y = mean(yy) %sum(yy)/n
var_x = var(xx,1) %sum(xx.^2)/n-moy_x^2
var_y = var(yy,1) %sum(yy.^2)/n-moy_y^2
cov_xy = cov(xx,yy,1) %sum(xx.*yy)/n-moy_x*moy_y
gamma_1 = cov_xy/var_x
gamma_0 = moy_y-gamma_1*moy_x
r_xy = cov_xy / sqrt(var_x*var_y)
```

Pour calculer la droite de régression de $\mathbf{x}$ par rapport à $\mathbf{y}$ on calcule les quantités suivantes :

$$C(\mathbf{y},\mathbf{x})=C(\mathbf{x},\mathbf{y})=\frac{21}{2}$$
$$\gamma_1'=\frac{C(\mathbf{y},\mathbf{x})}{V(\mathbf{y})}=\frac{3}{2}$$
$$\gamma_0'=\bar{\mathbf{x}}-\gamma_1'\bar{\mathbf{y}}=-\frac{1}{2}$$

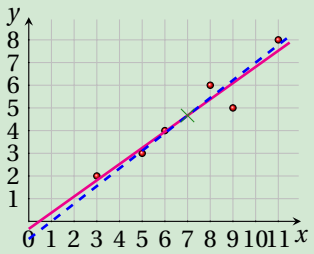
La droite cherchée a donc pour équation $x = \gamma'_0 + \gamma'_1 y = -\frac{1}{2} + \frac{3}{2}y$, soit encore $y = \frac{1}{3} + \frac{2}{3}x$.
On voit que $\gamma_1\gamma'_1 = \frac{7}{11}\frac{3}{2} = \frac{21}{22} = r^2$.

Exercice 1.4 (Régression linéaire)

Calculer les droites de meilleur approximation de l'ensemble de points suivant :

x	3	5	6	8	9	11
y	2	3	4	6	5	8

ainsi que leurs coefficients de corrélation.



Correction

Nous avons une série statistique double avec une population d'effectif $n = 6$.
Pour calculer la droite de régression de **y** par rapport à **x** on calcule les quantités suivantes :

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{42}{6} = 7$$
$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{28}{6} = \frac{14}{3}$$
$$V(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \bar{\mathbf{x}}^2 = \frac{336}{6} - 49 = 7$$
$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{\mathbf{x}}\bar{\mathbf{y}} = \frac{226}{6} - 7\frac{14}{3} = 5$$
$$\gamma_1 = \frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})} = \frac{5}{7}$$
$$\gamma_0 = \bar{\mathbf{y}} - \gamma_1\bar{\mathbf{x}} = -\frac{1}{3}$$
$$V(\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n} y_k^2 - \bar{\mathbf{y}}^2 = \frac{154}{6} - \frac{14^2}{9} = \frac{77\times 3 - 14^2}{9} = \frac{35}{9}$$
$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = \frac{3}{7}\sqrt{5} > 0.9$$

La droite cherchée a donc pour équation $y = \gamma_0 + \gamma_1 x = -\frac{1}{3} + \frac{5}{7}x$ avec une forte corrélation (mais cela ne dit rien sur la causalité entre les deux quantités!).

```
xx=[3,5,6,8,9,11]
yy=[2,3,4,6,5,8]

n = length(xx)
moy_x = mean(xx) %sum(xx)/n
moy_y = mean(yy) %sum(yy)/n
var_x = var(xx,1) %sum(xx.^2)/n-moy_x^2
var_y = var(yy,1) %sum(yy.^2)/n-moy_y^2
cov_xy = cov(xx,yy,1) %sum(xx.*yy)/n-moy_x*moy_y
gamma_1 = cov_xy/var_x
gamma_0 = moy_y-gamma_1*moy_x
r_xy = cov_xy / sqrt(var_x*var_y)
```

Pour calculer la droite de régression de **x** par rapport à **y** on calcule les quantités suivantes :

$$C(\mathbf{y},\mathbf{x}) = C(\mathbf{x},\mathbf{y}) = 5$$
$$\gamma'_1 = \frac{C(\mathbf{y},\mathbf{x})}{V(\mathbf{y})} = \frac{9}{7}$$
$$\gamma'_0 = \bar{\mathbf{x}} - \gamma'_1\bar{\mathbf{y}} = 1$$

La droite cherchée a donc pour équation $x = \gamma'_0 + \gamma'_1 y = 1 + \frac{9}{7}y$, soit encore $y = -\frac{7}{9} + \frac{7}{9}x$.
On voit que $\gamma_1\gamma'_1 = \frac{5}{7}\frac{9}{7} = \frac{45}{49} = r^2$.

Exercice 1.5

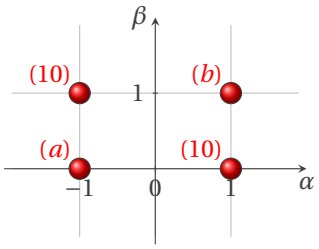
Soit le tableau de la distribution conjointe de deux variables quantitatives **x** et **y** :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$
$\alpha_1 = -1$ $\alpha_2 = 1$	$n_{1,1} = a$ $n_{2,1} = 10$	$n_{1,2} = 10$ $n_{2,2} = b$

1. Calculer les distributions marginales et écrire le tableau des fréquences de chaque couple et des fréquences marginales.
2. Calculer les distributions conditionnelles.
3. Calculer le coefficient de corrélation linéaire.

Correction

Ce tableau indique qu'on observe 10 fois le couple $(1,0)$, 10 fois le couple $(-1,1)$, a fois le couple $(-1,0)$ et b fois le couple $(1,1)$. On a donc au mieux $p \times q = 4$ points distincts (α_i, β_j) chacun avec un poids $n_{i,j}$:



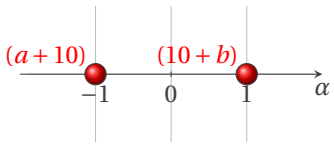
Si $a = b = 0$, alors on a seulement deux observations différentes sur deux variables (10 fois l'observation $(1,0)$ et 10 fois l'observation $(-1,1)$) : $r = -1$ (la droite de régression linéaire passe forcement par ces deux points et la pente est négative : la droite a pour équation $y = -\frac{1}{2}(x-1)$).
Si $a = b = 10$, il y a indépendance puisque les profils en lignes sont identiques donc $r = 0$ (la droite a pour équation $y = \frac{1}{2}$).
Si $a = 0$ et $b = 10$, il n'y a ni indépendance ($r \neq 0$), ni liaison linéaire ($r \neq \pm 1$). Même comportement si si $a = 10$ et $b = 0$.

Vérifions ce raisonnement par les calculs.

1. Distributions marginales
   - Effectifs marginaux de α_i :

     $$n_{1,\cdot} = 10 + a \qquad\qquad n_{2,\cdot} = 10 + b$$

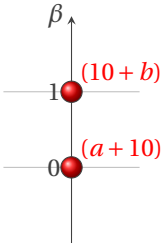
     et on a $\sum_{i=1}^{p=2} n_{i,\cdot} = 20 + a + b = n$. Autrement dit, indépendamment de l'observation de **y**, on observe $10 + a$ fois la valeur $x = \alpha_1 = -1$ et $10 + b$ fois la valeur $x = \alpha_2 = 1$.

   - Effectifs marginaux de β_j :

     $$n_{\cdot,1} = 10 + a \qquad\qquad n_{\cdot,2} = 10 + b$$

     et on a $\sum_{j=1}^{q=2} n_{\cdot,j} = 20 + a + b = n$. Autrement dit, indépendamment de l'observation de **x**, on observe $10 + a$ fois la valeur $y = \beta_1 = 0$ et $10 + b$ fois la valeur $y = \beta_2 = 1$.

- Tableau des effectifs

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	Effectif marginal de α_i
$\alpha_1 = -1$	$n_{1,1} = a$	$n_{1,2} = 10$	$n_{1,\cdot} = 10 + a$
$\alpha_2 = 1$	$n_{2,1} = 10$	$n_{2,2} = b$	$n_{2,\cdot} = 10 + b$
Effectif marginal de β_j	$n_{\cdot,1} = 10 + a$	$n_{\cdot,2} = 10 + b$	$n = 20 + a + b$

Tableau des fréquences

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	Fréquence marginale de α_i
$\alpha_1 = -1$	$f_{1,1} = \frac{a}{20+a+b}$	$f_{1,2} = \frac{10}{20+a+b}$	$f_{1,\cdot} = \frac{10+a}{20+a+b}$
$\alpha_2 = 1$	$f_{2,1} = \frac{10}{20+a+b}$	$f_{2,2} = \frac{b}{20+a+b}$	$f_{2,\cdot} = \frac{10+b}{20+a+b}$
Fréquence marginale de β_j	$f_{\cdot,1} = \frac{10+a}{20+a+b}$	$f_{\cdot,2} = \frac{10+b}{20+a+b}$	1

2. Distributions conditionnelles :
   - De **x** sachant **y** :
     - De **x** sachant β_1 (on ne regarde que la colonne $y = \beta_1$) :

$$f_{i=1|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{a}{10+a}, \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_1$$

$$f_{i=2|j=1} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{10}{10+a}, \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_1$$

$$\bar{x}_{j=1} = f_{i=1|j=1}\alpha_1 + f_{i=2|j=1}\alpha_2 = \frac{10-a}{10+a}$$

     - De **x** sachant β_2 (on ne regarde que la colonne $y = \beta_2$) :

$$f_{i=1|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{10}{10+b}, \qquad \text{fréquence conditionnelle de } \alpha_1 \text{ sachant } \beta_2$$

$$f_{i=2|j=2} = \frac{f_{i,j}}{f_{\cdot,j}} = \frac{b}{10+b}, \qquad \text{fréquence conditionnelle de } \alpha_2 \text{ sachant } \beta_2$$

$$\bar{x}_{j=2} = f_{i=1|j=2}\alpha_1 + f_{i=2|j=2}\alpha_2 = \frac{10-b}{10+b}$$

     - Tableau des profiles en colonne $f_{i|j}$:

Profiles en colonne $f_{i\|j}$		
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$
$\alpha_1 = -1$	$f_{1\|1} = \frac{a}{10+a}$	$f_{1\|2} = \frac{10}{10+b}$
$\alpha_2 = 1$	$f_{2\|1} = \frac{10}{10+a}$	$f_{2\|2} = \frac{b}{10+b}$
	1	1

   - De **y** sachant **x** :
     - De **y** sachant α_1 (on ne regarde que la ligne $x = \alpha_1$) :

$$f_{j=1|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{a}{10+a}, \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_1$$

$$f_{j=2|i=1} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{10}{10+a}, \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_1$$

$$\bar{y}_{i=1} = f_{j=1|i=1}\beta_1 + f_{j=2|i=1}\beta_2 = \frac{10}{10+a}$$

- De $\mathbf{y}$ sachant α_2 (on ne regarde que la ligne $x = \alpha_2$) :

$$f_{j=1|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{10}{10+b}, \qquad \text{fréquence conditionnelle de } \beta_1 \text{ sachant } \alpha_2$$

$$f_{j=2|i=2} = \frac{f_{i,j}}{f_{i,\cdot}} = \frac{b}{10+b}, \qquad \text{fréquence conditionnelle de } \beta_2 \text{ sachant } \alpha_2$$

$$\bar{y}_{i=2} = f_{j=1|i=2}\beta_1 + f_{j=2|i=2}\beta_2 = \frac{b}{10+b}$$

- Tableau des profiles en ligne $f_{j|i}$:

Profiles en ligne $f_{j\|i}$			
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 0$	$\beta_2 = 1$	
$\alpha_1 = -1$	$f_{1\|1} = \frac{a}{10+a}$	$f_{1\|2} = \frac{10}{10+a}$	1
$\alpha_2 = 1$	$f_{2\|1} = \frac{10}{10+b}$	$f_{2\|2} = \frac{b}{10+b}$	1

3. Calcul du coefficient de corrélation linéaire r :

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}\alpha_i = \frac{(10+a)\times(-1)+(10+b)\times(1)}{20+a+b} = \frac{b-a}{20+a+b},$$

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}\beta_j = \frac{(10+a)\times(0)+(10+b)\times(1)}{20+a+b} = \frac{10+b}{20+a+b},$$

$$V(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{p} n_{i,\cdot}\alpha_i^2 - \bar{\mathbf{x}}^2 = \frac{(10+a)\times(-1)^2+(10+b)\times(1)^2}{20+a+b} - \frac{(b-a)^2}{(20+a+b)^2} = 1 - \frac{(b-a)^2}{(20+a+b)^2} = 4\frac{ab+10a+10b+100}{(20+a+b)^2},$$

$$V(\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{q} n_{\cdot,j}\beta_i^2 - \bar{\mathbf{y}}^2 = \frac{(10+a)\times(0)^2+(10+b)\times(1)^2}{20+a+b} - \frac{(10+b)^2}{(20+a+b)^2} = \frac{10+b}{20+a+b}\left(1 - \frac{10+b}{20+a+b}\right) = \frac{(a+10)(b+10)}{(20+a+b)^2},$$

$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j}\alpha_i\beta_j - \bar{\mathbf{x}}\bar{\mathbf{y}} = \frac{a\times(-1)\times(0)+10\times(-1)\times(1)+10\times(1)\times(0)+b\times(1)\times(1)}{20+a+b} - \frac{(b-a)(10+b)}{(20+a+b)^2}$$

$$= 2\frac{ab-100}{(20+a+b)^2},$$

$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = \frac{ab-100}{\sqrt{a+10}\sqrt{b+10}\sqrt{ab+10a+10b+100}}.$$

En particulier on voit que si $a = b = 0$ alors $r = -1$, si $a = 0$ et $b = 10$ alors $r = -\frac{1}{2}$, si $a = b = 10$ alors $r = 0$.
De plus, si on calcule les coefficients γ_1 de la régression linéaire de $\mathbf{y}$ en fonction de $\mathbf{x}$ (pente de la droite) on trouve :

$$\gamma_1 = \frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})} = \frac{1}{2}\frac{ab-100}{ab+10a+10b+100}$$

Si $a = b$ alors $\gamma_1 = \frac{a-10}{2(a+10)}$: en particulier, si $a = b = 10$ alors $\gamma_1 = 0$. Si $a = 0$ alors $\gamma_1 = \frac{-5}{b+10} < 0$: en particulier, si $b = 0$ alors $\gamma_1 = -\frac{1}{2}$. Même calcul si $b = 0$ car $\gamma_1 = \frac{-5}{a+10} < 0$.

☆ Exercice 1.6
Reproduire les graphes de l'exemple à la page 1.13.

Correction

```
xx_A=[10.0 8.0 13.0 9.0 11.0 14.0 6.0 4.0 12.0 7.0 5.0];
yy_A=[8.04 6.95 7.58 8.81 8.33 9.96 7.24 4.26 10.84 4.82 5.68];
```

```
xx_B=[10.0 8.0 13.0 9.0 11.0 14.0 6.0 4.0 12.0 7.0 5.0];
yy_B=[9.14 8.14 8.74 8.77 9.26 8.10 6.13 3.10 9.13 7.26 4.74];
xx_C=[10.0 8.0 13.0 9.0 11.0 14.0 6.0 4.0 12.0 7.0 5.0];
yy_C=[7.46 6.77 12.74 7.11 7.81 8.84 6.08 5.39 8.15 6.42 5.73];
xx_D=[8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 19.0 8.0 8.0];
yy_D=[6.58 5.76 7.71 8.84 8.47 7.04 5.25 5.56 12.50 7.91 6.89];

T=[xx_A',yy_A',xx_B',yy_B',xx_C',yy_C',xx_D',yy_D'];
Cas=["A","B","C","D"];

for i=1:4
  CAS = Cas(i)
  xx = T(:,2*i-1);
  yy = T(:,2*i);
  r_xy = cov(xx,yy,1) / sqrt(var(xx,1)*var(yy,1))
  subplot(2,2,i)
  hold on
  plot(xx,yy,'*')
  gamma_1 = cov(xx,yy,1)/var(xx,1)
  gamma_0 = mean(yy)-gamma_1*mean(xx)
  d=@(x)gamma_0+gamma_1*x;
  plot(xx,d(xx),'r:')
  title(CAS)
  hold off
end
```

☆ Exercice 1.7
Reproduire les graphes de l'exemple à la page 1.14.

Correction

```
XX=[ 1.3 0.9 1 0.6 3 0.05 2 0.25 0.05 0.5 0.05 2.1 0.6 0.3 1.1 1.7 9 1.3 0.05 5 1.6];
YY=[439 650 150 34 651 11.5 230 7 11.5 434 31 255 35 385 228 383 1708 120 154 1172 3 ];

for i=1:2
  if i==1
    xx=XX;
    yy=YY;
  else % on enleve les deux valeurs extremes
    [val,idx]=max(YY);
    xx=XX([1:idx-1,idx+1:end]);
    yy=YY([1:idx-1,idx+1:end]);
    [val,idx]=max(yy);
    xx=xx([1:idx-1,idx+1:end]);
    yy=yy([1:idx-1,idx+1:end]);
  end
  figure()
  disp('')
  moy_x = mean(xx)
  moy_y = mean(yy)
  sigma_x = std(xx,1)
  sigma_y = std(yy,1)
  r_xy = cov(xx,yy,1) / sqrt(var(xx,1)*var(yy,1))
  hold on
  plot(xx,yy,'*')
  gamma_1 = cov(xx,yy,1)/var(xx,1)
  gamma_0 = mean(yy)-gamma_1*mean(xx)
  d=@(x)gamma_0+gamma_1*x;
  plot(xx,d(xx),'r:')
  hold off
end
```